

Pliego de Prescripciones Técnicas del contrato de servicios denominado “Ingesta de datos y explotación de la información en la plataforma de interoperabilidad de datos de la Dirección General de Estadística de la Comunidad de Madrid”

Objeto del contrato	2
Contenido del trabajo	3
Análisis funcional y diseño del modelo de datos	3
Normalización y estandarización que garantice la interoperabilidad sencilla de la información.....	3
Ingesta de información en la plataforma, tanto de fuentes que genera la Dirección General de Estadística de la Comunidad de Madrid, como fuentes de datos de terceros (otras administraciones, INE, Padrón, Catastro, IGN, etc.)	3
Transformación de la información en bruto para los distintos casos de uso de los datos en la plataforma y creación de tablas integradas a partir de distintas fuentes de información incorporadas a la plataforma.	4
Documentación en las herramientas de catalogación de datos, tanto de las fuentes de datos en bruto como de tablas derivadas en distintos procesos.	4
Explotación de la información a través de herramientas de cuadros de mando.....	4
Explotación de la información para la realización de informes ad-hoc, utilizando herramientas de análisis avanzado de datos, en entornos Data Science Open Source como R o Python.	4
Gestión del código que se genere y procesos de CI/CD	4
Gestionar la plataforma cloud de interoperabilidad de datos durante el período del servicio.....	5
Organización y seguimiento del proyecto	5
Gestión de las peticiones e incidencias de los distintos departamentos.	5
Composición del equipo del proyecto	5
Dependencia funcional del personal	8
Metodología de seguimiento del proyecto y comunicación con el personal involucrado de la Dirección General de Estadística.....	8
Recursos materiales del proyecto	8
Lugar de ejecución del proyecto.....	9
Plazo de ejecución	9
Presupuesto	9
Pago de la prestación	9
Secreto estadístico	9

Objeto del contrato

El objeto de este pliego es establecer las condiciones técnicas para la contratación, mediante procedimiento abierto, del contrato de servicios titulado: “Ingesta de datos y explotación de la información en la plataforma de interoperabilidad de datos de la Dirección General de Estadísticas de la Comunidad de Madrid”.

Tras la puesta en marcha de la plataforma de interoperabilidad de datos, es decir, los servicios tecnológicos y de infraestructura necesarios para soportar la carga de información y su posterior consumo, se plantea un contrato de servicios para operar dicha plataforma. En particular, las tareas que la empresa adjudicataria llevará a cabo serán las siguientes:

- Análisis funcional y diseño del modelo de datos
- Normalización y estandarización que garantice la interoperabilidad sencilla de la información
- Ingesta de información en la plataforma, tanto de fuentes que genera la Dirección General de Estadística de la Comunidad de Madrid, como fuentes de datos de terceros (otras administraciones, INE, Padrón, Catastro, IGN, etc.)
- Transformación de la información en bruto para los distintos casos de uso de los datos en la plataforma.
- Creación de tablas integradas a partir de distintas fuentes de información incorporadas a la plataforma.
- Documentación en las herramientas de catalogación de datos, tanto de las fuentes de datos en bruto como de tablas derivadas en distintos procesos.
- Explotación de la información a través de herramientas de cuadros de mando.
- Explotación de la información para la realización de informes ad-hoc, utilizando herramientas de análisis avanzado de datos, en entornos Data Science Open Source como R o Python.
- Gestión de las peticiones de los distintos departamentos.
- Gestión de las incidencias que puedan surgir en los procesos e información subida a la plataforma.
- Gestión del código que se genere y procesos de CI/CD
- Coordinar el seguimiento tanto con la Dirección General de Estadística de la Comunidad de Madrid como con los equipos involucrados en las distintas peticiones en ejecución.
- Gestionar la plataforma cloud de interoperabilidad de datos durante el período del servicio.

La empresa adjudicataria deberá dar solución y servicio a todas las necesidades de información de la Dirección General de Estadística de la Comunidad de Madrid, con experiencia en el ámbito de la gestión y gobernanza de los datos, así como en la explotación de los mismos. La empresa adjudicataria trabajará con los distintos departamentos de la dirección, canalizará sus peticiones y dará respuesta utilizando la plataforma de interoperabilidad de datos implantada, así como las herramientas de la misma.

Queda fuera del alcance de este proyecto cualquier coste en el que se pudiera incurrir por modificaciones de servicios en la plataforma de interoperabilidad de datos, así como ningún coste adicional de la infraestructura cloud.

Contenido del trabajo

Análisis funcional y diseño del modelo de datos

El equipo de la empresa adjudicataria involucrado en el proyecto será capaz de entender los requisitos de los distintos departamentos y traducirlos en especificaciones funcionales que puedan ser llevadas a cabo por los responsables de la ejecución de las tareas. Cuando estos requisitos impliquen modificar el modelo de datos existente, se notificará el impacto con otras fuentes y, en el caso de ingestión de una nueva fuente de datos, se relacionará con las tablas previas y se aplicarán estándares definidos tanto de formatos como semántica de los campos definidos.

En la propuesta es necesario que el proveedor defina a alto nivel cómo será el diseño de la arquitectura de la información, partiendo de la base:

- Habrá un nivel raw/staging para carga de datos en bruto directamente desde los orígenes
- Habrá uno o más niveles de capas de consumo, generales o en ámbitos concretos para aplicaciones o departamentos.
- Se debe tener en cuenta en el diseño el contexto concreto de la Dirección General de Estadística y se diferenciará en el modelo lógico, al menos, la información directamente generada propia o por terceros como aquella información generada para la elaboración de informes y estadísticas.

Normalización y estandarización que garantice la interoperabilidad sencilla de la información. De forma transversal al proyecto, se llevará a cabo un seguimiento de la estandarización y normalización de la información, esto es, la gestión de las reglas técnicas y semánticas que definen cada uno de los campos, que serán de obligatorio cumplimiento siempre que aparezca dicha información en una tabla. De esta manera, se asegurará la interoperabilidad futura de la información, se evitarán duplicidades y se generarán mecanismos de cruce simples entre las mismas.

En la propuesta es necesario que el proveedor defina las líneas generales de esta normalización y estandarización entre fuentes.

Ingesta de información en la plataforma, tanto de fuentes que genera la Dirección General de Estadística de la Comunidad de Madrid, como fuentes de datos de terceros (otras administraciones, INE, Padrón, Catastro, IGN, etc.)

Una carga importante del trabajo de este proyecto, sobre todo durante los primeros meses, será la ingesta de información en la plataforma de interoperabilidad de datos. En la propuesta enviada por el proveedor, deberá venir claramente especificado cómo se van a asegurar los siguientes puntos:

- Mecanismo de ingesta de la información en una capa inicial (raw o staging). Este punto debe considerar cómo se va a subir información a través de procesos automáticos, pero también procesos manuales donde un responsable de la fuente de datos suba la información a la plataforma. Se deberán explicar en la propuesta, al menos, ambos mecanismos.

- Mecanismos de transformación de la información para depositar el dato en capas de consumo (una capa Master, aunque podrá haber varias capas, según los usos de los datos).
- Generación de la metadata de cada una de las tablas ingestadas y generadas.
- Compatibilidad con el estándar SMDX.
- Procedimiento de documentación de los datos y mecanismos de Data Discovery internos en la organización.

Transformación de la información en bruto para los distintos casos de uso de los datos en la plataforma y creación de tablas integradas a partir de distintas fuentes de información incorporadas a la plataforma.

Una vez subida la información base a la plataforma de interoperabilidad de datos, muchas de las necesidades que tengan los distintos departamentos consistirán en el desarrollo de tablas derivadas que simplifiquen la posterior explotación de la información, generando nuevas variables a partir de cruces de tablas originales. Es necesario que el proveedor explique en qué tecnologías propone hacer este tipo de procesamientos, las ventajas e inconvenientes de cada uno y los perfiles profesionales, tanto propuestos por el proveedor como internos de la Dirección General de Estadísticas, que serán los consumidores o modificadores de dicha información.

Documentación en las herramientas de catalogación de datos, tanto de las fuentes de datos en bruto como de tablas derivadas en distintos procesos.

El proveedor deberá generar de manera sistemática y rigurosa la documentación de todas las tablas que se vayan subiendo o generando en la plataforma de interoperabilidad de datos, utilizando para ello las herramientas de catalogación de datos de las que disponga la Dirección.

Explotación de la información a través de herramientas de cuadros de mando.

El equipo propuesto por el proveedor deberá ser capaz de generar cuadros de mando utilizando para ello herramientas de Business Intelligence (Power BI, Tableau, Qlik View, etc.) así como herramientas de desarrollo en entornos típicos de Data Science (RShiny en R o Dash en Python). Los cuadros de mando deben tener la calidad suficiente para ser publicables a toda la ciudadanía y tienen que servir para apoyar los informes técnicos que la propia Dirección General de Estadística produce.

Explotación de la información para la realización de informes ad-hoc, utilizando herramientas de análisis avanzado de datos, en entornos Data Science Open Source como R o Python.

El equipo tendrá que ser capaz de realizar proyectos de analítica avanzada de datos, tanto análisis descriptivos más profundos a partir de distintas fuentes de información como, eventualmente, proyectos de Machine Learning e Inteligencia Artificial. Para ello, el equipo tendrá que dominar los ecosistemas de Data Science típicos basados en los lenguajes de programación R y Python. Se valorará, además, que el equipo tenga capacidades a la hora de realizar productivización de modelos, automatización de los procesos de generación de predicciones y desarrollo de APIs que permitan su integración en otros sistemas de información.

Gestión del código que se genere y procesos de CI/CD

El equipo aplicará las mejores prácticas del ámbito de desarrollo software, tanto para los procesos que se encarguen de ingestar la información, como para los procesos que generen

tablas derivadas o análisis más profundos. Esto implicará que, como mínimo, se utilizará una herramienta de gestionado de versión de código y, en la medida de lo posible, procesos integración y despliegue continuo sobre la arquitectura.

Gestionar la plataforma cloud de interoperabilidad de datos durante el período del servicio.

El proveedor será el encargado de gestionar la operación de la infraestructura y tecnologías de interoperabilidad de datos durante el servicio, esto es:

- Habilitar, modificar y eliminar usuarios en los distintos servicios.
- Conceder y eliminar accesos y permisos.
- Control del gasto de la infraestructura cloud.
- Modificación de servicios existentes.

Organización y seguimiento del proyecto

Gestión de las peticiones e incidencias de los distintos departamentos.

El equipo del proveedor será el encargado de gestionar las peticiones de los distintos departamentos apoyándose para ello en la herramienta de ticketing habilitada en la plataforma. Decidirá la prioridad de las peticiones de acuerdo con la Dirección de Estadística de la Comunidad de Madrid y será el encargado de transformar dichas peticiones en peticiones más concretas asignables a miembros del equipo que, eventualmente, informarán en la herramienta sobre el desarrollo y entregables de cada una de las tareas.

Composición del equipo del proyecto

En la propuesta, el proveedor deberá indicar:

- Perfiles profesionales asociados con tareas de gestión del proyecto y elaboración de entregables (figuras tipo Project Manager)
- Perfiles profesionales técnicos asociados con la ejecución de las tareas necesarias para llevar a cabo la implantación de la plataforma.

El proveedor adjuntará un Curriculum Vitae ciego para cada uno de los perfiles que indique en la propuesta. El proveedor se compromete a asegurar que el personal tiene las capacidades técnicas y conocimientos necesarios para la correcta ejecución del proyecto en tiempo y forma. Se valorará la experiencia individual de los miembros en proyectos similares.

Como guía general, como mínimo, el equipo deberá contar con los siguientes perfiles:

- Jefe de proyecto (1 recurso)
- Arquitecto de datos (1 recurso)
- Ingeniero de datos (3 recursos)
- Científicos de datos (2 recursos)

con los siguientes perfiles:

- DIRECTOR: Titulado de nivel postgrado con 5 años de experiencia demostrable en la dirección de proyectos de datos. Los recursos con este perfil deben cumplir los

siguientes requisitos:

- Titulaciones universitarias de grado y postgrado en ingeniería informática, matemáticas, ingeniería de telecomunicaciones, física, química o cualquier otra de carácter técnico-científico.
 - Experiencia contrastable en dirección proyectos de arquitectura e ingeniería de datos, pre-procesamiento de información, calidad de datos, etc. así como proyectos de analítica avanzada y modelos predictivos.
 - Experiencia contrastable en el uso metodologías de gestión de proyecto, tanto tradicionales (*waterfall*, *iterativas*, ...) como ágiles (Scrum, Agile, ...). Se valorará acreditaciones o certificaciones que demuestren, además de la experiencia, este conocimiento.
 - Experiencia en la dirección de proyectos basados en datos.
 - Experiencia en el uso de herramientas de gestión de proyectos como MS Project, Asana, Jira, Taiga, etc.
- **ARQUITECTO DE DATOS SENIOR:** Titulado postgrado con 5 años de experiencia demostrable en el diseño y ejecución de proyectos relacionados con la creación de repositorios de información. Los recursos con este perfil deben cumplir los siguientes requisitos:
 - Titulado de nivel postgrado en ingeniería informática, ingeniería de telecomunicaciones o cualquier otra de carácter tecnológico.
 - Experiencia demostrable en el análisis funcional, diseño de modelos de datos, generación de metadatos, uso del estándar SMDX en proyectos anteriores y desarrollo e implantación de metodologías para la gobernanza de los datos.
 - Experto en el desarrollo con lenguajes de programación Python, R, SQL y en el uso de herramientas ETL.
 - Experiencia demostrable en el desarrollo de tareas de ETL, Data Quality y pre-procesamiento en varios proyectos, dominando tanto el desarrollo como su puesta en producción.
 - Experiencia demostrable en el desarrollo de procesos batch y APIs, su orquestación, documentación automática y despliegue.
 - Experiencia demostrable en el uso Docker y Kubernetes para crear los entornos necesarios.
 - Experiencia demostrable en el uso de herramientas de control de versiones, crear pipelines para CI/CD, entornos virtuales, Jira, etc.
 - Capacidad de liderar reuniones con los equipos asignados a los proyectos.
 - Capacidad de generar entregables de calidad: documentación, código y presentaciones.

- **INGENIERO DE DATOS:** Titulado de grado 2 años de experiencia demostrable en la ejecución de proyectos relacionados con la creación de repositorios de información. Los recursos con este perfil deben cumplir los siguientes requisitos:
 - Titulaciones universitarias de grado en ingeniería informática, ingeniería de telecomunicaciones o cualquier otra de carácter tecnológico.
 - Experto en el desarrollo con lenguajes de programación Python, R, SQL y en el uso de herramientas ETL.
 - Experiencia demostrable en el desarrollo de tareas de ETL, Data Quality y pre-procesamiento en varios proyectos, dominando tanto el desarrollo como su puesta en producción.
 - Experiencia demostrable en el desarrollo de procesos batch y APIs, su orquestación, documentación automática y despliegue.

- **CIENTÍFICO DE DATOS:** Titulado de grado 2 años de experiencia demostrable en la ejecución de proyectos relacionados con el análisis de información con técnicas de aprendizaje automático y machine learning. Los recursos con este perfil deben cumplir los siguientes requisitos:
 - Titulaciones universitarias de grado en ingeniería informática, ingeniería de telecomunicaciones, matemáticas, estadística, física o cualquier otra de carácter tecnológico.
 - Experto en el uso de los lenguajes de programación Python, R y SQL.
 - Experiencia en el uso de herramientas de visualización de datos como Power BI, Tableau, RShiny o Dash.
 - Experiencia en el desarrollo tareas de ETL, Data Quality y pre-procesamiento en varios proyectos, dominando tanto el desarrollo como su puesta en producción.
 - Experiencia en la realización tareas de análisis descriptivo y de negocio en distintos proyectos, realizando interpretaciones y recomendaciones sobre los resultados obtenidos.
 - Experto en el desarrollo de procesos de generación de variables y feature engineering, así como técnicas de reducción de la dimensionalidad previas al entrenamiento de modelos.
 - Experto en el desarrollo de algoritmos de Aprendizaje Automático para tareas de clasificación, segmentación, regresión, procesamiento de lenguaje natural y algoritmos de visión artificial
 - Aplica best practices de desarrollo: herramientas de control de versiones, crear pipelines para CI/CD, entornos virtuales, Jira, etc.

- Capacidad de liderar reuniones con los equipos asignados a los proyectos.
- Capacidad de generar entregables de calidad: documentación, código y presentaciones.

Además, como norma general, para el equipo en general, se valorará la experiencia demostrable en proyectos de similares características al del objeto del presente contrato.

El proveedor se compromete a no variar el equipo destinado al proyecto salvo causas no controlables como el posible abandono del personal de la empresa.

La Dirección General de Estadística de la Comunidad de Madrid podrá solicitar el cambio de miembros del equipo con 15 días de antelación, lo que siempre será justificado por razones relacionadas con el inadecuado desarrollo de la prestación del servicio.

En caso de sustitución del personal técnico, los candidatos deberán mantener un perfil de cualificación técnica igual o superior al de la persona o personas sustituidas.

Los perfiles de ingenieros y científicos de datos estarán dedicados 100% al presente proyecto mientras que el proveedor puede indicar dedicaciones parciales para el jefe de proyecto y el arquitecto de datos, justificando en todo caso los motivos de dicha dedicación.

Dependencia funcional del personal

La empresa adjudicataria deberá aportar el personal preciso para cumplir las obligaciones derivadas de la ejecución de este contrato. Dicho personal dependerá exclusivamente del adjudicatario, por cuanto este tendrá todos los derechos y deberes inherentes a su calidad de empresario y deberá cumplir con las disposiciones vigentes en materia laboral, de seguridad social y de seguridad e higiene en el trabajo, sin que en ningún caso pueda alegarse derecho alguno por dicho personal en relación con la Administración contratante, ni exigirse a esta responsabilidades de cualquier clase como consecuencia de las obligaciones existentes entre el adjudicatario y sus empleados, aún en el supuesto de que los despidos o medidas que se adopte se basen en el incumplimiento, interpretación o resolución de este contrato.

Metodología de seguimiento del proyecto y comunicación con el personal involucrado de la Dirección General de Estadística

Para asegurar el correcto desarrollo de acuerdo con las condiciones de las Prescripciones Técnicas, este se llevará a cabo bajo el control y supervisión del Subdirector General de Estadística de la Dirección General de Estadística y, a quien el Director de Proyecto designado por el adjudicatario informará del estado del proyecto.

Adicionalmente, el proveedor propondrá una metodología de seguimiento con el personal técnico involucrado en el proyecto y un modelo de comunicación a distintos niveles, con el fin de garantizar la agilidad operativa del proyecto y el reporte de los resultados parciales y totales del mismo.

Recursos materiales del proyecto

El proveedor se compromete a proporcionar al equipo destinado todos los medios materiales necesarios para el correcto desarrollo del proyecto.

El proveedor, como mínimo, tendrá que poner a disposición del equipo del proyecto:

- Mobiliario de oficina
- Equipos informáticos
- Conexión a internet
- Herramientas software necesarias

Y en general cualquier otro medio material que sea necesario para la correcta ejecución del contrato

Lugar de ejecución del proyecto

Si bien el proyecto se podrá ejecutar de manera remota desde las oficinas del proveedor, realizando el seguimiento de forma telemática o presencial a requerimiento de la Dirección General de Estadística de la Comunidad de Madrid, dado que mucha de la información del proyecto reside en instancias locales de la Dirección General, es necesario que el equipo pueda desplazarse a las oficinas de la Dirección General de Estadística de la Comunidad de Madrid con total flexibilidad para trabajar en ellas de forma presencial. El proveedor no podrá repercutir costes adicionales derivados de estas solicitudes de trabajo presencial.

Plazo de ejecución

36 meses

Presupuesto

El presupuesto total del proyecto es de 1.647.310,92 euros, impuestos incluidos.

Pago de la prestación

El pago de la prestación se realizará de forma semestral. En la facturación, el proveedor indicará las tareas objeto del contrato realizadas por los miembros del equipo. Los pagos parciales se realizarán tras las correspondientes certificaciones de las tareas realizadas durante el periodo correspondiente.

Secreto estadístico

La empresa adjudicataria queda sujeta al secreto estadístico en lo que respecta a toda la información que pase por sus manos por motivo de esta operación estadística (Ley de Estadística de la Comunidad de Madrid; art. 15 y siguientes), con las obligaciones que se derivan para todo el personal. Todo el personal implicado en los trabajos deberá conocer estas circunstancias y firmar un documento en el que afirme conocer y tener voluntad de cumplir estos extremos. Dichos documentos serán entregados a la Dirección General de Estadística.

En Madrid, a fecha de la firma

LA DIRECTORA GENERAL DE ESTADÍSTICA

Firmado digitalmente por: DE LA FUENTE CORRALES MARÍA
Fecha: 2020.12.02 14:36