

Dossier técnico

Inteligencia artificial para la gestión municipal

Este documento describe, con detalle técnico, cómo funciona Noctis Public: la arquitectura del sistema, cada una de sus funciones, la metodología con la que se ha verificado su fiabilidad, las medidas de seguridad que incorpora y el proceso de implantación en un ayuntamiento. Está pensado para que un equipo técnico, un secretario-interventor o un regidor puedan evaluar el sistema con la misma seriedad con la que evaluarían cualquier otro proveedor.

El problema que resuelve

La atención ciudadana en la mayoría de ayuntamientos depende de canales que no escalan: una línea telefónica con horario limitado, un mostrador presencial, o un WhatsApp que alguien del equipo revisa cuando puede. El resultado es previsible: preguntas repetidas que consumen tiempo de personal cualificado, ciudadanos que no saben qué documentación llevar a una cita, incidencias en la vía pública que se pierden entre llamadas, y una sensación general de que la administración va un paso por detrás de lo que la ciudadanía espera de cualquier servicio digital hoy en día.

Noctis Public no resuelve esto añadiendo otro canal más que vigilar. Resuelve esto gestionando el trámite completo -- desde la intención inicial del ciudadano hasta la resolución o la derivación a una persona -- con un principio de diseño que se aplica sin excepciones: el sistema nunca simula certeza que no tiene. Cuando no dispone de un dato verificado, lo dice explícitamente y deriva. Esa regla, más que cualquier función concreta, es lo que hace que el sistema sea utilizable en un contexto donde un error de información tiene consecuencias reales para un ciudadano.

Contenido de este documento

1. Arquitectura técnica
2. Funciones del sistema, explicadas en detalle
3. Metodología de verificación y resultados
4. Seguridad y cumplimiento normativo
5. Casos de uso ilustrativos
6. Proceso de implantación
7. Preguntas frecuentes

95,8%

de acierto en 500 casos de prueba con consultas reales

4

capas de seguridad activas desde el primer mensaje: firma, hash, TLS, auditoría

100%

de las respuestas trazables: cada acción queda registrada de forma inmutable

1. Arquitectura técnica

El sistema está construido como un backend propio en Python (FastAPI), no como un flujo dentro de una herramienta de automatización genérica. Esa decisión de diseño tiene consecuencias concretas en seguridad, trazabilidad y control que se explican a continuación.

Capa de entrada: verificación criptográfica del canal

Cada mensaje entrante -- ya llegue por WhatsApp Business API o por el canal web -- pasa primero por una verificación de firma HMAC-SHA256. Esto garantiza que la petición procede realmente del proveedor del canal y no de un tercero que intenta inyectar mensajes falsos en el sistema. Es una capa de seguridad que muchas integraciones genéricas (flujos montados sobre herramientas tipo Make o Zapier) no implementan por defecto.

Motor de clasificación y respuesta

El texto del mensaje se envía a un modelo de lenguaje (Claude, de Anthropic) para clasificar la intención del ciudadano -- reservar, cancelar, reportar una incidencia, consultar información -- y generar la respuesta. Por principio de minimización de datos, no se envían identificadores directos del ciudadano al motor de IA salvo que el propio ciudadano los escriba voluntariamente en su mensaje. El modelo no tiene acceso de escritura a ninguna base de datos: solo clasifica y redacta; la ejecución de la acción (reservar, cancelar, registrar) la realiza código determinista, no el modelo.

Persistencia y auditoría

Los datos se almacenan en una base de datos Postgres alojada en la Unión Europea. Además de las tablas operativas (citas, incidencias, catálogo de trámites), existe una tabla de auditoría de solo inserción: el rol de aplicación no tiene permisos de modificación ni borrado sobre ella. Cada acción relevante -- una cita reservada, una incidencia registrada, una derivación a personal -- queda escrita ahí de forma permanente, con marca temporal.

El identificador del ciudadano (su número de WhatsApp) nunca se almacena en claro en los registros de auditoría. Se guarda como una huella irreversible generada mediante función hash SHA-256 con una clave de aplicación (pepper), de forma que no es posible reconstruir el número original a partir del registro, pero sí es posible confirmar si dos entradas corresponden al mismo ciudadano cuando haga falta, por ejemplo ante una auditoría o una solicitud de ejercicio de derechos.

Gestión de secretos y entornos

Todos los secretos -- tokens de acceso, claves de API, credenciales de base de datos -- se gestionan mediante variables de entorno o un gestor de secretos, nunca escritos directamente en el código ni versionados en el repositorio. Los entornos de desarrollo y producción están aislados, con credenciales distintas para cada uno, evitando que un fallo en pruebas pueda afectar a datos reales.

Integraciones externas

El sistema se apoya en un número reducido de proveedores externos, cada uno con una función acotada: el canal de mensajería (Meta, para WhatsApp Business API), el motor de inteligencia artificial (Anthropic), la sincronización de calendario (Google Calendar) y el hosting de base de datos (infraestructura en la UE). Ningún proveedor externo tiene acceso a más datos de los estrictamente necesarios para la función que presta.

2. Funciones del sistema, explicadas en detalle

Cada función descrita a continuación está implementada como un módulo independiente del sistema, no como una simulación de lo que la IA cree que debería responder.

2.1 Gestión de citas presenciales

El sistema consulta la disponibilidad real en el calendario del departamento correspondiente antes de proponer una hora -- nunca ofrece un hueco sin comprobarlo primero. Gestiona reserva, modificación y cancelación de forma autónoma, respetando reglas de duración por tipo de trámite y capacidad del departamento. Al confirmar la cita, genera automáticamente un PDF de checklist con la documentación exacta que el ciudadano debe llevar, extraída del catálogo verificado de ese ayuntamiento -- no de una respuesta genérica producida por el modelo de lenguaje.

Ejemplo de flujo: un ciudadano escribe pidiendo cita para renovar el DNI. El sistema identifica el trámite, consulta el calendario del departamento de atención ciudadana, propone tres horas disponibles esta semana, confirma la elegida, y envía el PDF con la documentación necesaria (DNI anterior, una fotografía reciente, justificante de pago de tasa). Si el ciudadano pregunta después si puede cambiar la hora, el sistema reconoce la cita existente y gestiona el cambio sin que tenga que empezar de nuevo.

2.2 Incidencias en la vía pública

El ciudadano puede reportar una incidencia (una farola averiada, un socavón, mobiliario urbano dañado) adjuntando foto y, si lo autoriza, ubicación. La incidencia se registra con un número de referencia rastreable y se notifica automáticamente al personal responsable del área correspondiente. Un detalle de diseño relevante en términos de privacidad: la fotografía nunca se descarga ni se almacena en la infraestructura del sistema -- se guarda únicamente el identificador del archivo en la plataforma de mensajería, que el personal municipal consulta directamente ahí cuando lo necesita.

Ejemplo de flujo: un vecino envía una foto de una farola apagada junto con su ubicación. El sistema crea la incidencia con referencia INC-004821, la clasifica en el área de alumbrado público, notifica al departamento correspondiente, y responde al ciudadano confirmando el número de referencia para que pueda hacer seguimiento si vuelve a escribir preguntando por el estado.

2.3 Modo consulta para personal municipal

Un modo de acceso restringido a números de teléfono autorizados por el propio ayuntamiento, pensado para que el personal consulte las citas e incidencias del día sin necesidad de acceder a un panel web independiente. Este modo es deliberadamente de solo lectura: no permite modificar, cancelar ni crear nada, reduciendo el riesgo de que un error de manejo desde el móvil altere datos por accidente.

2.4 Checklists de documentación en PDF

Cada trámite del catálogo tiene asociada su documentación exacta. Cuando se confirma una cita o se consulta información sobre un trámite, el sistema genera al vuelo un PDF con esa checklist, evitando que el ciudadano acuda sin lo necesario y tenga que volver otro día -- una de las causas

más frecuentes de ineficiencia en la atención presencial.

2.5 Verificación de fuentes oficiales

El sistema compara periódicamente el contenido de las páginas oficiales del ayuntamiento (requisitos de trámites, horarios, tasas) contra el catálogo interno que utiliza para responder. Cuando detecta una diferencia, genera una alerta para revisión humana. Es una decisión de diseño deliberada que el sistema nunca actualice el catálogo por su cuenta: la verificación automática reduce el trabajo de mantenimiento, pero la decisión de qué información es correcta la toma siempre una persona.

2.6 Auditoría de solo inserción

Descrita con más detalle en la sección de arquitectura: cada acción del sistema queda registrada de forma permanente e inmodificable. Esto permite, ante cualquier duda o revisión posterior, reconstruir exactamente qué respondió el sistema, cuándo, y con base en qué información -- una capacidad de trazabilidad que un canal de atención humana informal (por ejemplo, WhatsApp gestionado manualmente por un funcionario) no ofrece de forma nativa.

3. Metodología de verificación y resultados

Antes de considerar el sistema listo para una implantación real, se sometió a una batería de 500 casos de prueba contruidos para representar cómo escribe la ciudadanía de verdad: mensajes ambiguos, con faltas, formulados de varias maneras distintas para la misma intención, y casos límite donde la respuesta correcta es reconocer que falta información.

Qué se probó

- Reservas, modificaciones y cancelaciones de citas, incluyendo peticiones ambiguas o incompletas.
- Reportes de incidencias con distintos niveles de detalle en la descripción.
- Consultas de información general sobre trámites y requisitos.
- Casos diseñados específicamente para comprobar que el sistema deriva a una persona o admite no saber algo, en lugar de inventar una respuesta plausible.

Resultado y tratamiento de los fallos

El sistema resolvió correctamente el 95,8% de los 500 casos (479 de 500). Cada uno de los fallos restantes se documentó y analizó de forma individual -- no se descartaron como ruido estadístico. Ese análisis caso por caso es el que ha ido informando ajustes posteriores en la clasificación de intención y en los criterios de derivación a personal humano.

Es importante ser precisos sobre qué mide esta cifra: es la precisión del sistema sobre un conjunto de pruebas diseñado para ser representativo, no una garantía de que el 95,8% de las interacciones reales con la ciudadanía de un ayuntamiento concreto se resolverán sin intervención humana -- ese porcentaje real depende del volumen y la naturaleza de los trámites de cada municipio, y se calibra durante la fase de pruebas controladas de la implantación (ver sección 6).

4. Seguridad y cumplimiento normativo

Protección de datos (RGPD / LOPDGDD)

La implantación se formaliza con un contrato de encargo de tratamiento entre el ayuntamiento (responsable del tratamiento) y Noctis Systems (encargado del tratamiento), conforme al artículo 28 del RGPD y a la LOPDGDD. Ese contrato identifica de forma explícita los subencargados que participan en la prestación del servicio (proveedor del canal de mensajería, proveedor de IA, hosting de base de datos, sincronización de calendario), las medidas de seguridad aplicadas, y el destino de los datos al finalizar la relación contractual.

El sistema aplica minimización de datos por diseño: solo se trata la información estrictamente necesaria para gestionar el trámite. Los archivos multimedia de incidencias no se descargan ni almacenan. Los identificadores de ciudadanos se seudonimizan en los registros de auditoría.

Esquema Nacional de Seguridad (ENS)

El sistema incorpora medidas de seguridad alineadas con los principios del ENS desde su diseño: verificación criptográfica de las comunicaciones, cifrado en tránsito, seudonimización de identificadores, y un registro de auditoría inmutable. La categoría ENS formalmente aplicable (básica, media o alta) depende de la naturaleza de los datos que trate cada ayuntamiento en su implantación concreta, y se determina conjuntamente con el propio ayuntamiento como parte del proceso de configuración inicial, de forma que la certificación se ajuste exactamente al uso real que se le va a dar al sistema.

Resumen de medidas técnicas activas

- Verificación de firma HMAC-SHA256 en cada mensaje entrante al webhook.
- Cifrado en tránsito (TLS) en todas las comunicaciones con proveedores externos.
- Seudonimización irreversible de identificadores de ciudadanos en los registros de auditoría.
- Registro de auditoría de solo inserción, sin permisos de modificación ni borrado para la aplicación.
- No descarga ni almacenamiento de archivos multimedia de incidencias.
- Gestión de secretos mediante variables de entorno / gestor de secretos, nunca en el código fuente.
- Separación de entornos de desarrollo y producción, con credenciales independientes.

5. Casos de uso ilustrativos

Los siguientes escenarios son ejemplos ilustrativos de cómo opera el sistema en la práctica, contruidos a partir de las funciones reales descritas en este documento -- no son testimonios de un cliente concreto.

Escenario 1 — Fuera del horario de atención

Un vecino escribe un domingo por la noche preguntando por el horario de recogida de muebles. Ninguna oficina está abierta. El sistema consulta el catálogo verificado, confirma el horario y el procedimiento (dónde y cuándo dejar los enseres), y ofrece programar un aviso si el ciudadano lo necesita. Nadie del ayuntamiento ha tenido que intervenir, y el ciudadano ha recibido una respuesta correcta sin esperar al lunes.

Escenario 2 — Una situación que el sistema no puede resolver solo

Un ciudadano escribe una queja formal sobre la actuación de un agente de la policía local durante un control de tráfico. El sistema reconoce que esto excede lo que puede gestionar de forma automática -- no es una consulta de información ni un trámite estándar -- y deriva la conversación al personal correspondiente, dejando registrado todo el contexto previo para que quien atienda no tenga que pedir que se repita la situación desde cero.

Escenario 3 — Personal municipal consultando el día

Un funcionario del departamento de vía pública, desde su propio teléfono, escribe para consultar las incidencias abiertas ese día. Al estar su número autorizado en modo consulta, el sistema le devuelve el listado con referencias y ubicaciones, sin que pueda modificar ni cerrar ninguna desde ahí -- esa acción requiere el canal de gestión correspondiente, no el chat.

6. Proceso de implantación

La implantación no consiste en activar el sistema y esperar. Sigue cuatro fases, cada una con un resultado verificable antes de pasar a la siguiente.

Fase 1 · Diagnóstico y catálogo verificado

Se revisan con el ayuntamiento los trámites, citas e incidencias que gestiona actualmente, y se construye el catálogo de información verificada -- requisitos, documentación, horarios, tasas -- que el sistema usará como única fuente de verdad. Ningún dato entra en el catálogo sin validación por parte de personal del ayuntamiento.

Fase 2 · Configuración conjunta

Se definen los idiomas activos, el canal o canales de despliegue (WhatsApp, web institucional, o ambos), los números de personal autorizados para el modo consulta, y las reglas exactas de cuándo el sistema debe derivar a una persona en lugar de responder de forma automática.

Fase 3 · Pruebas controladas antes de producción

El sistema se somete a una batería de pruebas adaptada a los trámites reales de ese municipio concreto -- del mismo tipo que la descrita en la sección 3, pero con el catálogo específico del ayuntamiento -- antes de exponerse a un solo ciudadano real. El ayuntamiento revisa los resultados antes de dar luz verde al lanzamiento.

Fase 4 · Verificación continua tras el lanzamiento

Una vez en producción, el sistema sigue comprobando periódicamente si las fuentes oficiales del ayuntamiento han cambiado (ver sección 2.5), y genera alertas para revisión humana. La relación no termina en el lanzamiento: el sistema requiere mantenimiento del catálogo igual que cualquier canal de información oficial.

7. Preguntas frecuentes

¿Sustituye al personal del ayuntamiento?

No. Gestiona el trabajo repetitivo -- reservar, informar, generar documentación -- y deriva a una persona cuando la situación lo requiere de verdad. El personal se centra en lo que necesita criterio humano, no en responder la misma pregunta cien veces.

¿Solo funciona por WhatsApp?

No. El motor de gestión es independiente del canal. Puede desplegarse por WhatsApp, integrarse en la web institucional del ayuntamiento, o en ambos a la vez.

¿Qué pasa si el sistema no sabe algo?

Lo dice explícitamente y deriva a personal municipal. Es una regla de diseño aplicada de forma consistente en todos los módulos: nunca simula una respuesta que no puede confirmar.

¿Puede consultar el estado de un expediente o una multa?

Hoy no. Esa función depende de integrarse con el sistema de gestión de expedientes propio de cada ayuntamiento, algo que se evalúa caso a caso según la infraestructura existente.

¿Qué ocurre con las fotografías de las incidencias?

Nunca se descargan ni se almacenan en la infraestructura del sistema. Se referencia únicamente el identificador del archivo en la plataforma de mensajería.

¿Cómo se decide la categoría ENS aplicable?

Se determina conjuntamente con el ayuntamiento durante la fase de configuración, en función de la naturaleza de los datos que se vayan a tratar en esa implantación concreta.

¿Cuánto tarda una implantación completa?

Depende del volumen de trámites a catalogar y de la disponibilidad del personal municipal para validar el catálogo inicial. El proceso no se acelera saltándose la fase de pruebas controladas.

¿Hablamos de cómo encaja en tu ayuntamiento?

noctissystems.com/noctis-public.html