# BASE DE DATOS DE ACTIVIDAD (BDA): UNA APLICACIÓN PARA LA GESTIÓN Y EL TRATAMIENTO DE REGISTROS ADMINISTRATIVOS DE ACTIVIDAD Y MERCADO LABORAL

Marta Mas Moreno m-mas@eustat.eus

Nerea Linaza Diarce nerea_linaza@eustat.eus

Instituto Vasco de Estadística

(Eustat)

Desde 2008, con la creación de la sub-área Base de Datos Sociodemográficas (BSD), ha aumentado de forma considerable el número de ficheros administrativos con información sobre actividad y mercado laboral que se reciben de forma periódica en el Eustat procedentes de diversos Organismos: Tesorería General e Instituto Nacional de la Seguridad Social, LANBIDE, MUFACE, Diputaciones Forales y Departamento de Educación. El gran volumen de datos y la frecuencia con la que se reciben, además de la necesidad de utilización de estos ficheros por parte de diversas operaciones estadísticas dentro del Instituto, hacen necesaria la gestión y administración diferenciada de esta información dentro del Sistema de Información de Eustat.

Para facilitar esta gestión se pretende poner en marcha un nuevo repositorio de almacenamiento y consulta de fuentes administrativas que permitirá poner a disposición de terceros dentro del Instituto la información administrativa sobre actividad y mercado laboral. El proyecto también incluirá el diseño y desarrollo de una aplicación informática 'ad-hoc' que facilite la ejecución de los tratamientos específicos necesarios para cada fuente de la manera más automatizada posible. Se trata de que dichas fuentes estén listas en tiempo y con la calidad suficiente para ser utilizadas por las operaciones estadísticas y recursos del Eustat que así lo requieran.

Actualmente, el grado de automatización de los tratamientos implementados en muchas de las fuentes permite plantear un tipo de aplicación que realice "llamadas" a procesos ya programados en otros entornos (Módulo de lectura de ficheros, Módulo de fusión, etc.). Además esta aplicación permitirá el desarrollo y mejora de los procesos de actualización y mantenimiento de las Unidades de Población Básica (UPB's) que funcionan como identificador único e intransferible de las personas registradas y como clave interna entre los distintos sub-sistemas de información de Eustat.

Palabras clave: *registro administrativo, relación con la actividad, mercado de trabajo, base de datos*.

## 1. Antecedentes y justificación

Desde 2008, con la creación de la sub-área Base de Datos Sociodemográficas (BSD), ha aumentado de forma considerable el número de ficheros administrativos con información sobre actividad y mercado laboral que se reciben de forma periódica en el Eustat procedentes de diversos Organismos: Tesorería General e Instituto Nacional de la Seguridad Social (TGSS e INSS), Servicio vasco de empleo (Lanbide), Mutualidad General de funcionarios civiles del Estado (MUFACE), Diputaciones Forales y diferentes departamentos del Gobierno Vasco. El gran volumen de datos y la frecuencia con la que se reciben, además de la necesidad de utilización de estos ficheros por parte de distintas operaciones estadísticas dentro del Instituto, hacen necesaria la gestión y administración diferenciada de esta información dentro del Sistema de Información de Eustat.

Para facilitar esta gestión se pretende poner en marcha un nuevo repositorio de almacenamiento y consulta de fuentes administrativas que permitirá poner a disposición de terceros dentro del Instituto la información administrativa sobre actividad y mercado laboral. El proyecto también incluirá el diseño y desarrollo de una aplicación informática 'ad-hoc' que facilite la ejecución de los tratamientos específicos necesarios para cada fuente de la manera más automatizada posible. Se trata de que dichas fuentes estén listas en tiempo y con la calidad suficiente para ser utilizadas por las operaciones estadísticas y recursos del Eustat que así lo requieran.

Además esta aplicación permitirá el desarrollo y mejora de los procesos de actualización y mantenimiento de las Unidades de Población Básica (UPB's) que funcionan como identificador único e intransferible de las personas registradas y como clave interna entre los distintos sub-sistemas de información de Eustat.

## 2. Gestión de las fuentes administrativas en Eustat

Cada vez más, un mayor número de operaciones en Eustat basan total o parcialmente su recogida de datos en información procedente de registros administrativos. Concretamente, el uso de registros administrativos sobre mercado laboral y relación con la actividad ha adquirido una enorme dimensión dentro del Instituto lo que ha planteado la necesidad de implantar una gestión y tratamiento diferenciados del resto de fuentes.

Operaciones estadísticas tales como la Estadística municipal de población activa (EMPA), la Encuesta de población en relación con la actividad (PRA) o el Directorio de actividades económicas (DIRAE) acceden periódicamente a estas fuentes para alcanzar sus objetivos de calidad en cuanto a la producción y difusión de información sobre actividad y mercado laboral.

Estas fuentes son principalmente:

- Fichero de perceptores de rentas reflejados en el modelo anual de retenciones (modelo 190) y proporcionado por las Haciendas Forales de los tres Territorios Históricos.

- Ficheros de afiliación y cuentas de cotización de la Seguridad Social enviados por Tesorería General de la Seguridad Social.
- Ficheros de demandantes de empleo y contratos realizados facilitados por Lanbide.
- Fichero de titulares de pensiones contributivas que proporciona el INSS.
- Fichero de mutualistas vascos que constan como tal en MUFACE.

Además estas son actualmente las principales operaciones estadísticas usuarias de registros administrativos de actividad en Eustat:

- Estadística municipal de población activa
- Encuesta de población en relación con la actividad
- Directorio de actividades económicas
- Estadística municipal de educación (EME)
- Estadística de población y viviendas
- Censos de población y viviendas

Actualmente, el protocolo aplicado cuando llega cualquiera de las fuentes administrativas utilizadas en Eustat sigue el siguiente esquema:

- *Llegada de los ficheros:* Se reciben los ficheros a través de sistemas seguros de intercambio de información. Los sistemas comúnmente utilizados en el Instituto son EDITRANS, cliente IFI y HARTU. Estos sistemas generan alarmas de aviso que informan a los usuarios interesados de la llegada de los ficheros y mantienen los registros de entradas y salidas de los mismos.

- *Lectura, generación y tratamiento de* ***tablas temporales*** según los diseños predefinidos para cada fuente. Gran parte de los ficheros originales son leídos a través de un módulo de carga de ficheros (CARFI). Se trata de una aplicación diseñada 'ad-hoc' para la lectura de ficheros de grandes dimensiones y está preparada para ejecutar los tratamientos iniciales de normalización y depuración de datos más habituales.

- *Generación de las* ***tablas definitivas***. También a través de CARFI se aplican tratamientos específicos a las tablas temporales ajustados a las características y peculiaridades de cada fuente. Estos tratamientos se clasifican generalmente en:
    - *Codificación de campos concretos y generación de características*.
    - *Asignación de claves mediante* f*usión con registro de población.* Una de las prioridades de Eustat consiste en el mantenimiento de una clave interna única e individual que permita la fusión entre las distintas fuentes y ficheros que se manejan en el Instituto. Para los registros referidos a personas esta clave individual es la UPB (Unidad de Población Básica) y su asignación y mantenimiento corresponde al Registro de población de Eustat. Por lo tanto, es necesario fusionar con este Registro de población cada fuente nueva que llega al Instituto. Esta fusión se realiza a través del Módulo de fusión (MDF), de nuevo una aplicación diseñada 'ad-hoc' para este

cometido. El mantenimiento de ficheros históricos de las distintas fuentes administrativas permite volcar las UPB's de periodos anteriores en el último periodo recibido y fusionar únicamente aquellos registros "nuevos" que aparecen en cada fichero.

- *Tratamiento y actualización de las claves identificativas.* Los cambios en la población de referencia del registro (altas, bajas, modificaciones), así como la unificación de algunas claves incorrectamente asignadas al mismo individuo, hace que los ficheros administrativos requieran de una actualización y revisión periódica de sus claves identificativas. Estos procesos de actualización y revisión de claves se realizan actualmente a través de la aplicación CARFI.
- *Carga de variables padronales del registro de población.* Los campos de situación de la clave identificativa unidos a las variables geográficas asociadas a ésta, y que constan en el registro de población, permiten realizar controles sobre el número de registros identificados en cada fuente y tras cada actualización de las mismas.

Tras estos tratamientos se generan las tablas definitivas que contendrán los campos originales y generados que necesita la operación estadística concreta para sus fines, así como la clave de identificación única. Además, estas tablas, como norma general, no contendrán identificadores directos de las personas en aras de preservar el secreto estadístico tal y como consta en el apartado 7 del art. 20 de la Ley 4/1986, de 16 de abril, de Estadística de la Comunidad Autónoma de Euskadi.

El siguiente diagrama resume el actual proceso de gestión de fuentes administrativas en Eustat:

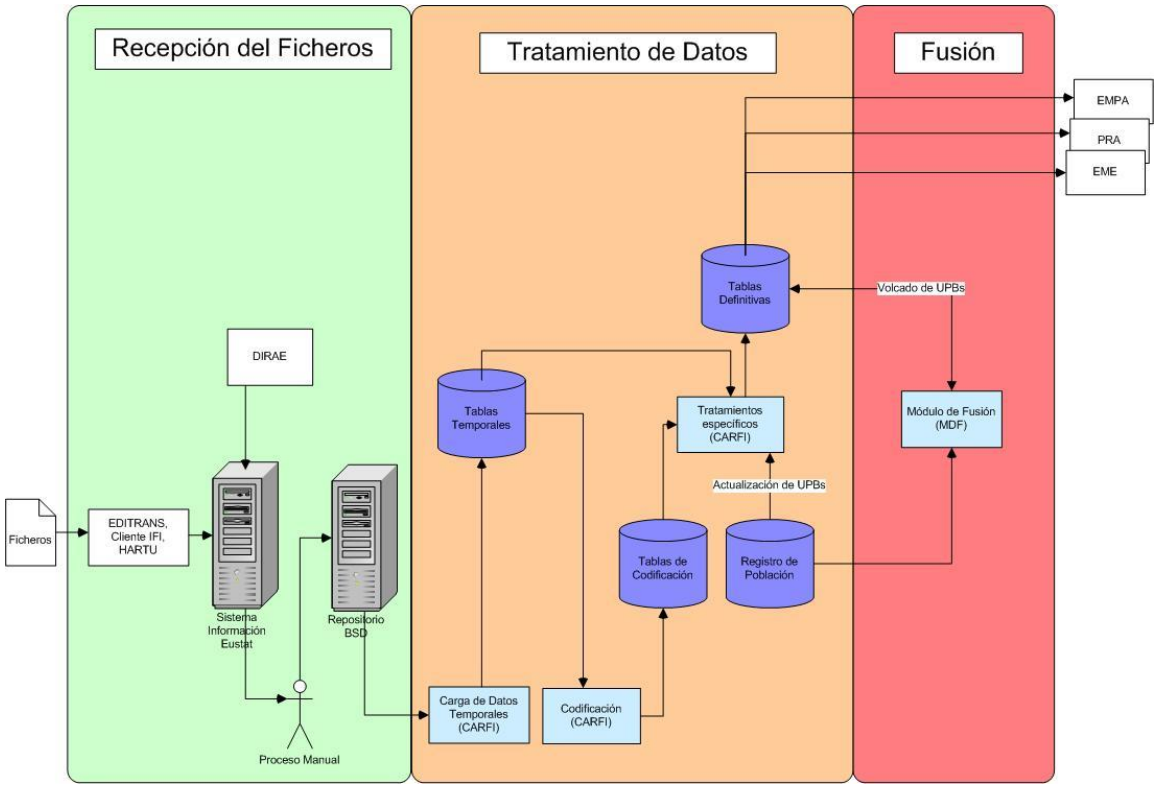


Figura 1

Estos pasos se realizan normalmente de forma secuencial, para cada fuente y periodo, por los usuarios autorizados que acceden, en cada caso, a los distintos módulos y programas que ejecutan los tratamientos descritos.

Además, cuando es necesario, está previsto el acceso a determinadas fuentes administrativas por parte de los usuarios de Eustat que así lo soliciten, ya sea a las tablas temporales (con objeto de ejecutar sobre éstas tratamientos específicos de cada operación) o a las definitivas, aprovechando de esa forma los tratamientos a los que éstas ya han sido sometidos. Este acceso se proporciona de forma puntual o temporal a la fuente solicitada y siempre que el fin esté justificado y sea proporcional.

Esta forma de trabajo, tanto por el volumen de datos que maneja como por la diversidad de programas y módulos que requiere para su ejecución, puede resultar, en ocasiones, muy laboriosa y prolongada en el tiempo, por lo que se requiere de un sistema que simplifique las tareas de gestión de los ficheros y las tablas y mejore la eficiencia de los tratamientos aplicados.

## 3. Gestión de las fuentes administrativas en la nueva BDA

La aplicación de Base de Datos de Actividad (BDA) debe contemplar las opciones necesarias para aplicar el protocolo de recepción y tratamiento de ficheros descrito en el apartado anterior, así como proporcionar un espacio común de consulta para aquellos usuarios y operaciones de Eustat que requieran el acceso a las fuentes administrativas de actividad.

Actualmente, el grado de automatización de los tratamientos aplicados a muchas de las fuentes permite plantear un tipo de aplicación que realice “llamadas” a procesos ya programados en otros entornos (CARFI, MDF, etc.).

El siguiente diagrama muestra el esquema funcional de la BDA, mucho más amigable que el actual, basado en el acceso automático a los diferentes entornos y programas:

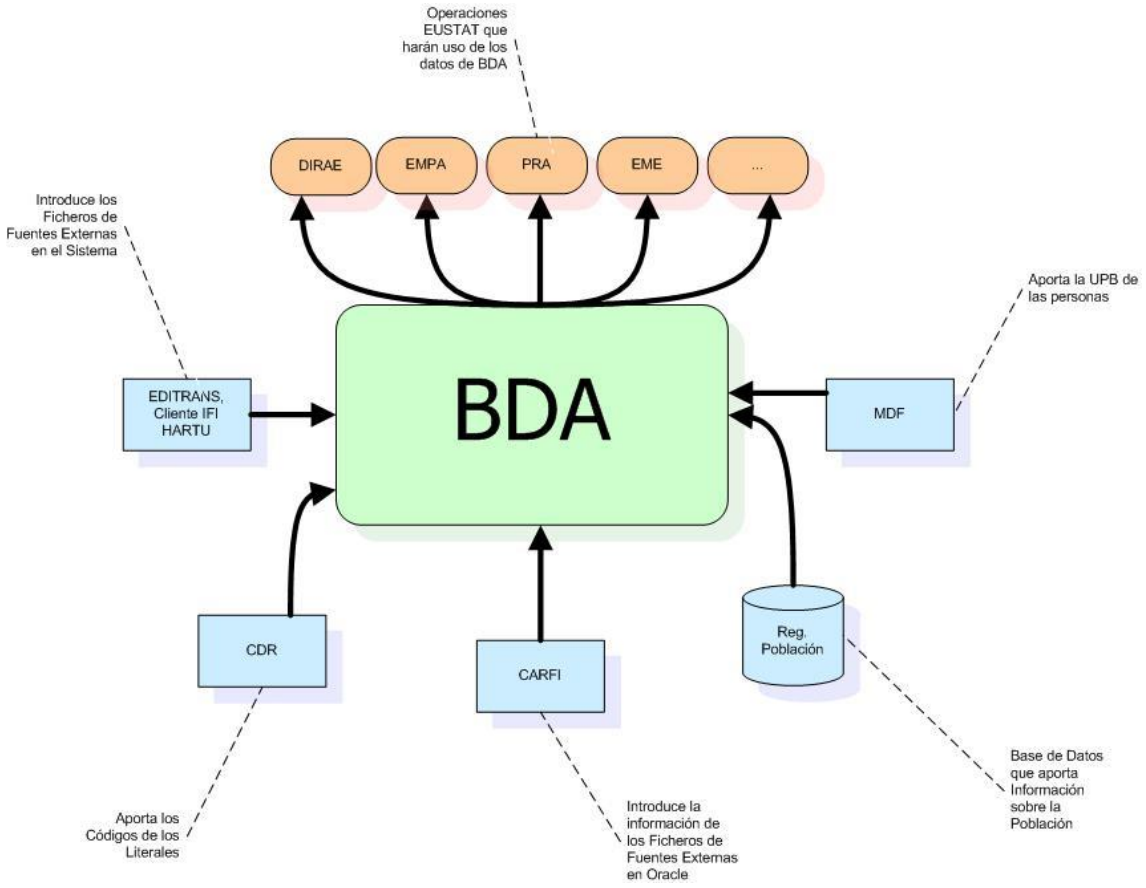


Figura 2

Para implementar este esquema, se han analizado los distintos tratamientos que se aplican en el caso de una fuente trimestral tipo (ficheros de cuentas y afiliados de la Seguridad Social) y otra anual (fichero de retenciones de las diputaciones forales). En general, se establecen tres grandes bloques de tratamientos que integran todos los procesos que se llevan a cabo actualmente desde que se reciben los ficheros fuente hasta que están listos para ser utilizados.

La idea es que cada bloque pueda ser lanzado como una única unidad desde la aplicación BDA. Además será necesario fijar los parámetros que requiera cada bloque para su ejecución, la periodicidad y los procesos o llamadas a módulos externos que se deberán realizar en cada caso.

## Ficheros trimestrales de cuentas y afiliados de la Seguridad Social

### Bloque 1: Generación del fichero temporal

- Lectura de los ficheros de texto de cuentas de cotización y trabajadores.
- Generación de la tabla temporal según diseño de registro.

*Parámetros:* Ubicación ficheros de texto
Año y trimestre de referencia de los datos (AAAAT)

*Periodicidad:* Trimestral

*Fecha de ejecución:* Mes 1 del trimestre T+1

*Procesos implicados:* Módulo de carga de ficheros (CARFI)

### Bloque 2: Generación del fichero trimestral definitivo

- Selección de registros individuales según criterios de prioridad preestablecidos
- Depuración y codificación de variables, codificación de literales de calles (llamada a CDR)
- Volcado de las claves identificativas del periodo anterior
- Generación de variables – Situación profesional (1ªParte)
- Fusión con el registro de población – Llamada al Módulo de Fusión (MDF)
- Volcado de nuevas claves identificativas encontradas
- Generación de variables - Situación profesional (2ª Parte)
- Tratamiento de claves identificativas
    - Paso de claves a periodos anteriores
    - Actualizar claves unificadas
    - Eliminar claves duplicadas
- Carga y actualización de variables padronales (de situación y geográficas)

*Parámetros:* Año y trimestre de referencia de los datos (AAAAT)
Año de referencia del registro de población (AAAA-1, donde AAAA es el año de referencia de los datos)
Nº de periodos (para el paso de UPBS' a periodos anteriores)

*Periodicidad:* Trimestral

*Fecha de ejecución:* Mes 1 del trimestre T+1

*Procesos implicados:* Módulo de carga de ficheros (CARFI), Módulo de fusión (MDF)

### Bloque 3: Fusión de los ficheros trimestrales con la nueva población de referencia

- Fusión con el registro de población más actualizado
- Volcado de nuevas claves identificativas encontradas
- Tratamiento de claves identificativas
    - Paso de claves identificativas a periodos anteriores
    - Actualizar claves unificadas
    - Eliminar claves duplicadas
- Carga y actualización de variables padronales (de situación y geográficas)

*Parámetros:* Año y trimestres de referencia de los datos (AAAAT)
Año de referencia del registro de población (AAAA ó AAAA+1)
Nº de periodos (para el paso de UPBS' a periodos anteriores)

*Periodicidad:* Anual

*Fechas de ejecución:* Enero del AAAA+1 (contra población AAAA)
Enero del AAAA+2 (contra población AAAA+1)

*Procesos implicados:* Módulo de carga de ficheros (CARFI), Módulo de fusión (MDF)

**Cuadro resumen tratamientos:**

| Fichero | Tratamiento | Period. | Fecha ejecución | Proceso CARFI | Proceso MDF |
|---|---|---|---|---|---|
| BSD_SSOC_AAAAT | Bloque 1: Generación del fichero temporal. | Trimestral | Mes 1 del trimestre T+1 | ✓ | |
| BSD_SSOC_AAAAT | Bloque 2: Generación del fichero trimestral definitivo. | Trimestral | Mes 1 del trimestre T+1 *(incluye fusión contra población AAAA-1)* | ✓ | ✓ |
| BSD_SSOC_AAAAT | Bloque 3: Fusión con nueva población de referencia. | Anual | Enero del AAAA+1** *(contra población AAAA)*<br><br>Enero del AAAA+2 *(contra población AAAA+1)* | ✓ | ✓ |

* AAAA se refiere al año de referencia de los datos y T al trimestre.
** Tras publicación de PRA último trimestre y media anual.

## Fichero anual de retenciones Modelo 190 - RETEN

### Bloque 1: Generación de los ficheros temporales

- Lectura de los ficheros de texto de retenciones (uno por cada TH).
- Generación de las tablas temporales (una por cada TH).

*Parámetros:* Ubicación ficheros de texto
Año de referencia de los datos (AAAA)

*Periodicidad:* Anual

*Fecha de ejecución:* 1er mes 2º trimestre del año AAAA+1

*Procesos implicados:* CARFI

### Bloque 2: Generación del fichero anual definitivo

- Carga secuencial de los ficheros temporales de cada territorio
- Depuración y codificación de variables
- Volcado de claves identificativas del periodo anterior
- Fusión con el registro de población
- Volcado de nuevas claves encontradas
- Tratamiento de claves identificativas
    - Paso de claves a periodos anteriores
    - Actualizar claves unificadas
    - Eliminar claves duplicadas
- Carga y actualización de variables padronales (de situación y geográficas)

*Parámetros:* Año de referencia de los datos (AAAA)
Año de referencia del registro de población (AAAA)
Nº de periodos (para el paso de UPBS' a periodos anteriores)
*Periodicidad:* Anual

*Fecha de ejecución:* 1er mes 2º trimestre del año AAAA+1

*Procesos implicados:* CARFI, MDF

**Bloque 3: Fusión del fichero anual con la nueva población de referencia**

- Fusión con el registro de población más actualizado
- Volcado de nuevas claves identificativas encontradas
- Tratamiento de claves identificativas
    - Paso de claves a periodos anteriores
    - Actualizar claves unificadas
    - Eliminar claves duplicadas
- Carga y actualización de variables padronales (de situación y geográficas)

*Parámetros:* Año de referencia de los datos (AAAA)
Año de referencia del registro de población (AAAA+1)
Nº de periodos (para el paso de UPBS' a periodos anteriores)

*Periodicidad:* Anual

*Fecha de ejecución:* Enero del AAAA+2 (contra población AAAA+1)

*Procesos implicados:* CARFI, MDF

## Cuadro resumen tratamientos:

| Fichero | Tratamiento | Period. | Fecha ejecución | Proceso CARFI | Proceso MDF |
|---|---|---|---|---|---|
| BSD_RETEN_AAAA | Bloque 1: Generación de los ficheros temporales. | Anual | 1er mes 2º trimestre del año | ✓ | |
| BSD_RETEN_AAAA | Bloque 2: Generación del fichero anual definitivo. | Anual | 1er mes 2º trimestre del año *(incluye fusión contra población AAAA)* | ✓ | ✓ |
| BSD_RETEN_AAAA | Bloque 3: Fusión con nueva población de referencia. | Anual | Enero del AAAA+2 *(contra población AAAA+1)* | ✓ | ✓ |

* AAAA se refiere al año de referencia de los datos.

Para ejecutar cada uno de los bloques de tratamientos y para cada fuente será necesario establecer una relación de usuarios autorizados. Además se debería de contar con un panel de control o una pantalla de estado que permita conocer a cualquier usuario autorizado en qué fase del tratamiento se encuentra cada fuente, fechas y horas de ejecución de los tratamientos, usuario que lo ha ejecutado, etc.

## 4. Actualización y mantenimiento de las claves identificativas

Además de la funcionalidad principal que permitirá ejecutar los tratamientos de forma automática para cada fuente, se plantean otras cuestiones que han venido afectando al análisis de las fuentes antes y después de cada tratamiento, sobre todo en lo que se refiere a la actualización de las claves identificativas cuando cambia la población de referencia.

Generalmente el desfase de la población respecto al periodo de referencia de las fuentes administrativas suele ser de dos años ya que para el periodo de referencia de datos del año AAAA se suele disponer de la población a 1 de enero del año AAAA-1, lo que a los efectos de la actualización de la población corresponde al periodo AAAA-2. Por lo tanto, para mantener actualizadas las claves identificativas en los ficheros conforme al periodo de referencia de los datos, habrá que ir actualizando dichas claves en cada periodo de actualización del registro de población. Esto supone un mínimo de tres actualizaciones por fichero, hasta que el periodo de referencia de los datos coincida con el de la población de referencia.

Estas actualizaciones, que se llevan a cabo para todas las fuentes administrativas, conllevan un trabajo periódico con una gran carga computacional. De cara a agilizar estos trabajos y mejorar el control de estas actualizaciones está prevista la introducción de una serie de mejoras que se incluirán en la BDA y que se resumen a continuación:

- *Depurar el tratamiento de duplicados de claves identificativas:* Existe la posibilidad de eliminar la clave duplicada manteniendo el registro duplicado o eliminar todo el registro (actualmente se está eliminando el registro completo, lo que da lugar a discrepancias en el total de registros antes y después de la actualización de claves).

- *Depurar el tratamiento de claves unificadas:* El registro de población unifica claves distintas que, de forma incorrecta, se han asignado a la misma persona. Estas duplicidades se corrigen eligiendo una de entre las distintas claves asignadas a la misma persona (clave unificada). Es necesario trasladar estas actualizaciones a los ficheros fuente y sobretodo determinar hasta qué periodos de referencia actualizar.

- *Contemplar un módulo específico de análisis de claves* que permita analizar y describir los cambios de claves identificativas en cada fuente cuando cambia la población de referencia. (P.ej.: nº de claves unificadas, actualizadas, duplicadas, eliminadas, relación de claves afectadas por estos cambios, etc.).

## 5. Consulta y almacenamiento de fuentes

La aplicación como herramienta de consulta y almacenamiento de fuentes administrativas permitirá el acceso a las fuentes existentes a los usuarios autorizados y proporcionará una información mínima sobre su contenido, periodicidad y disponibilidad.

Además la aplicación debe integrar y mejorar el actual sistema de “alertas” con el Centro de Proceso de Datos (CPD) para la recogida de información. Este sistema informará al CPD de las fechas clave de recepción de ficheros y, a su vez, remitirá los avisos pertinentes a los responsables del tratamiento de la información cuando ésta llegue.

## 6. Otras aplicaciones de la BDA

La implantación de la BDA puede facilitar el acceso a las fuentes administrativas de actividad y mercado laboral por parte de otras operaciones estadísticas y mejorar su uso por parte de las que ya las utilizan actualmente. Posibles aplicaciones de este entorno pueden ser las siguientes:

- DIRAE, indicadores económicos, estadísticas coyunturales y, en general, estadísticas basadas en establecimientos económicos para la depuración del empresariado, personas autónomas y cooperativistas, mejora en la recogida del empleo, inclusión de la perspectiva de género, volcado de claves identificativas de población para las personas físicas, etc.
- Estadística Universitaria y Estadística Municipal de Educación para completar itinerarios profesionales, integración laboral según nivel de estudios, etc.
- Estadística de nacimientos, matrimonios y defunciones para completar y actualizar la información sobre profesión y relación con la actividad de los progenitores, cónyuges, etc.
- Registro de población para depurar y corregir información identificativa, sobretodo de la población extranjera (DNI’s, nombres, apellidos, fechas de nacimiento, etc.).
- Enlace con otros módulos y sistemas: Normalización y homogeneización de direcciones, territorio, etc.

Por último, y aunque la actividad y el mercado laboral generan actualmente el volumen más importante de información administrativa en el Eustat, es importante realizar una reflexión sobre el creciente uso de fuentes administrativas en todos los ámbitos temáticos que abarca la estadística oficial (p.ej.: ficheros catastrales, universidades, etc.). Un repositorio de estas características supondría solo el punto de partida para la creación de una *base de datos administrativos compartida* que proporcionaría el acceso a estas fuentes a todos los usuarios y operaciones del Instituto que así lo requieran.

## 7. Planificación y tareas previstas.

Está previsto que los trabajos de diseño e implantación de la BDA finalicen en el primer trimestre del próximo año. Actualmente se ha llevado a cabo el análisis de viabilidad, conjuntamente con el área de informática, y es inminente el comienzo de los trabajos de desarrollo de la aplicación. Durante esta fase se pretende llevar a cabo pruebas piloto sobre las fuentes administrativas “tipo” y reuniones de seguimiento que involucrarán al resto de las áreas del Instituto, potenciales usuarias de la aplicación.