IT ACTIVITIES ORIENTED TO THE DATA PROCESS OPTIMIZATION IN THE SPANISH LFS

Information and Communication Technologies Directorate
Jorge Velasco
May 2012

Background

- The deadline of delivery of the quarterly LFS file to Eurostat is 12 weeks after the end of the reference period.

- Spanish National Statistics Institute (INE) publishes the results of the Spanish LFS a month after the end of the quarter
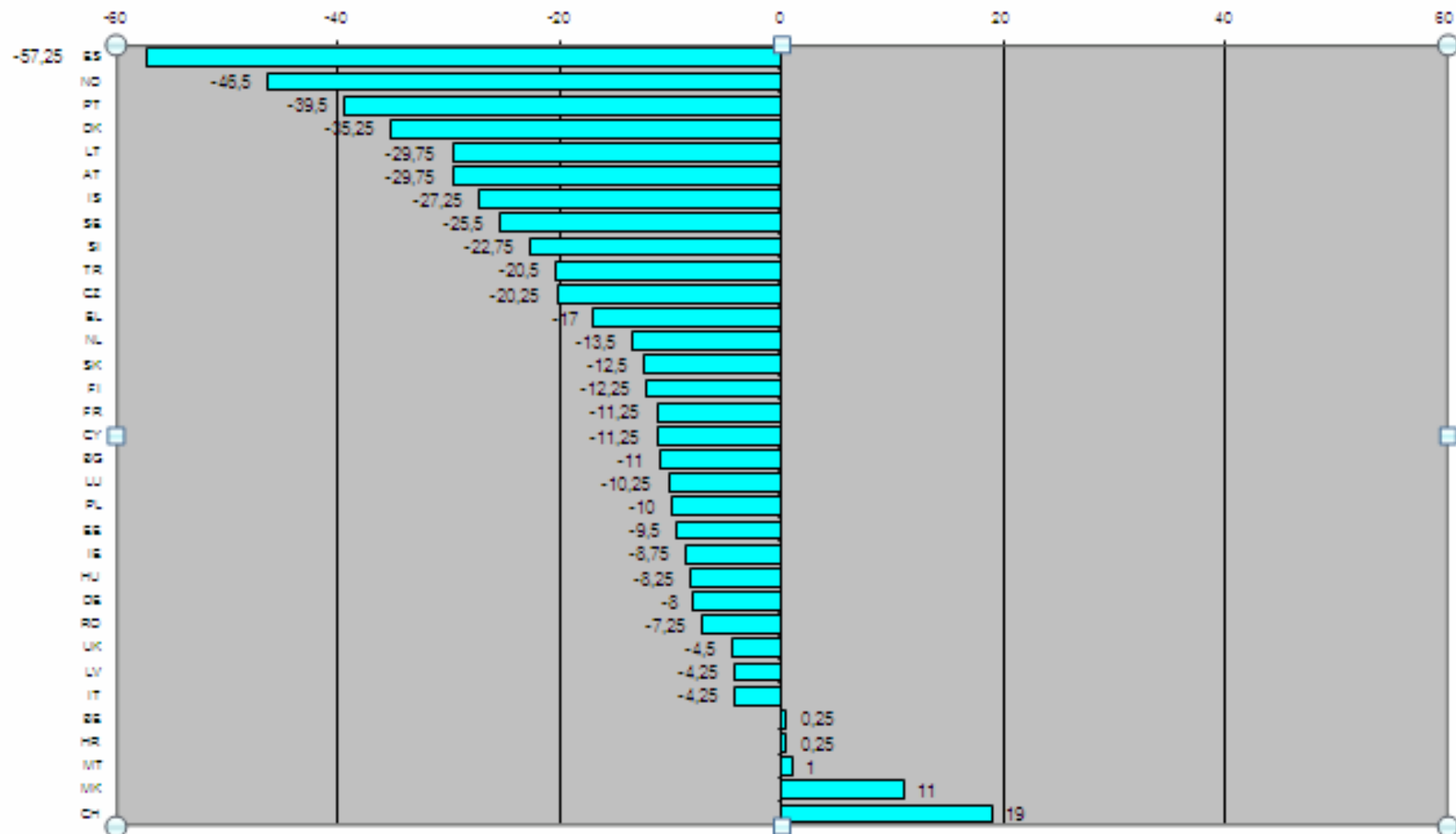
Timeliness

2010 quarterly LFS-average. Files accepted for dissemination.

Key points

■ There are several factors that contribute to this short time of data dissemination, and the **optimization of the technological processes** is one amongst them:

◆ Commitment to do it

◆ Timing and organization of the fieldwork to ensure a suitable pace in data collection

◆ Availability of population figures for weighting in time

◆ ...

◆ **Optimization of the technological processes**

Optimization of the technological processes

■ At the ***data collection stage***

◆ → Different factors and actions from the technological viewpoint that allow the **early disposal** with quality **of the data** at the next stage in the survey process, the data process

◆ → Leads to data process optimization

◆ → Leads to short time of data dissemination

5

Optimization of the technological processes

- If the data is collected with the required quality and on time, the continuity of the survey process transfers to having available a contrasted and fault tolerant **Data Process Stage**

Management of survey processes

Sample Preparation
Workgroup Preparation
Calendar (in weeks)
Workplan (in workgrups-weeks)

Preparing the cycle

Management of the collection

Letters and labels
Workgroup management
Assignation of CAPI work batch
Reserve dwellings
Transfer of dwellings CAPI-CATI
Call control system

Fieldwork

Interviews (CAPI-CATI)
Data Recording
Inspections
Fieldwork notes
Cover and Backcover
Appointments

Monitoring of the collection

Listeners
900 line
Tables and monitoring reports
partial results linstings

closure of cicle and distribution of results

closure of sections and dwellings
cycle rotation
data publication

Centralized Treatment

Host data transfer
Debugging
Imputation

Dump of data from Delegations to Headquarters

Factors and activities at the ***data collection stage***:

◆ 2.1.Using an electronic questionnaire

◆ 2.2.Technological Processes

◆ 2.3.Effective field work application. Organization and monitoring of field work

◆ 2.4.Weekly download

◆ 2.5.Collection and data delivery schedule

- Factors and activities at the ***data collection stage***:

- ◆ **<u>2.1.Using an electronic questionnaire</u>**

- ◆ 2.2.Technological Processes

- ◆ 2.3.Effective field work application. Organization and monitoring of field work

- ◆ 2.4.Weekly download

- ◆ 2.5.Collection and data delivery schedule

2.1.Using an electronic questionnaire

- CAPI method for the first interview and CATI for second and subsequent

- The use of an electronic questionnaire ensures the quality of the information collected, because it includes online rules for inconsistency and flow validations while collection.

- →data received for data process are already pre-depurated

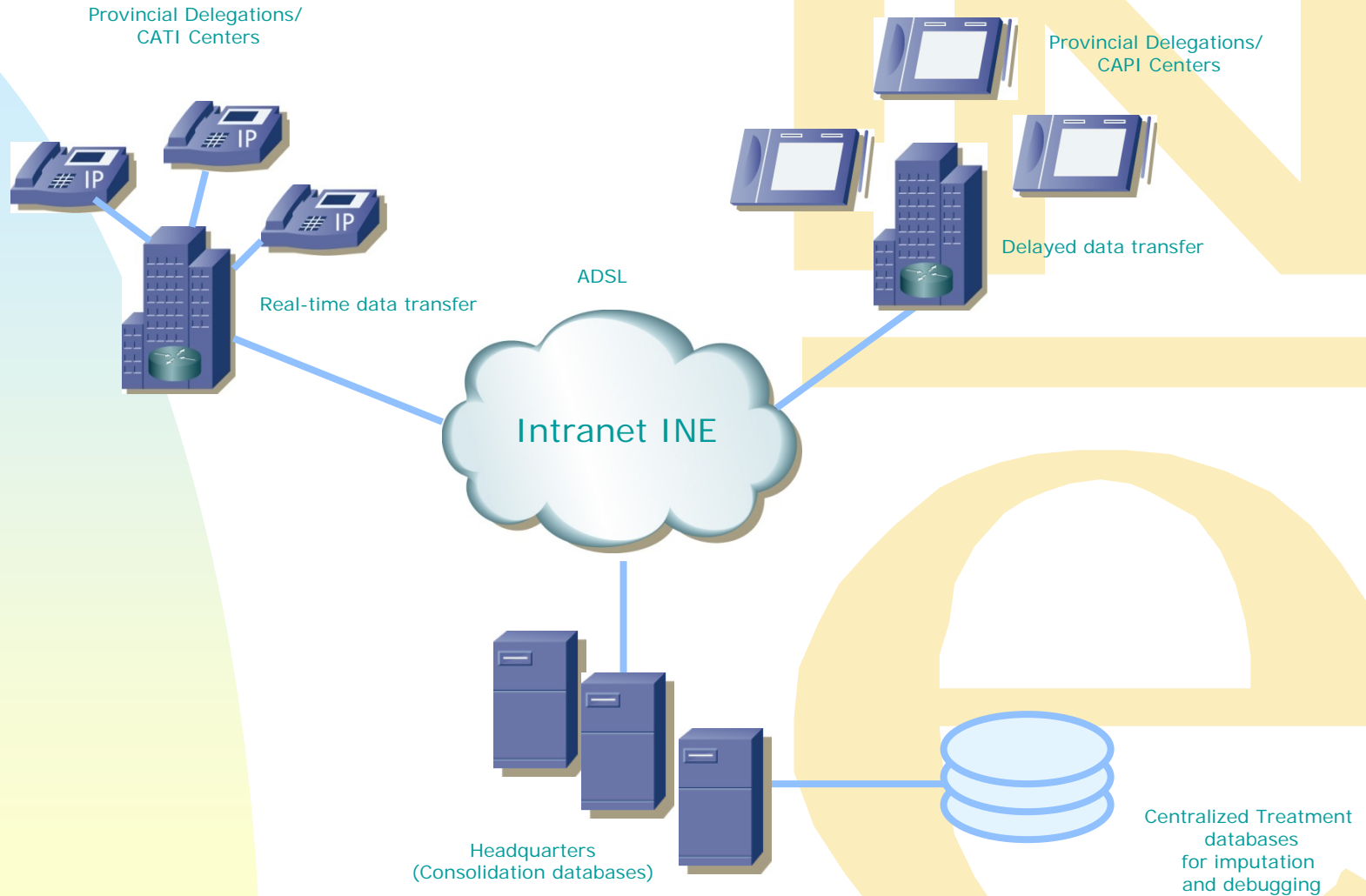- →reduces data process time as there are fewer errors to be debugged

Collecting flow data

Provincial Delegations/
CATI Centers

Provincial Delegations/
CAPI Centers

1. Background

**2. Data
Collection
Stage**

3. Data Process
Stage

Delayed data transfer

ADSL

Real-time data transfer

Intranet INE

Headquarters
(Consolidation databases)

Centralized Treatment
databases
for imputation
and debugging

Personas de 16 y más años

A. DATOS GENERALES

1. Número de la persona: _____   **NPERS**

2. Nombre y apellidos: _____   **NOMBRE/ APELLIDOS**

3. Entrevistador, anote quién responde a este cuestionario: **PROXY**
   - La propia persona   1 ☐
   - Otra persona de la vivienda de 16 y más años   2 ☐
   - Negativa a responder   3 ☐   Fin de encuesta

4. De las personas relacionadas anteriormente, ¿alguna de ellas es su
   (Codificar 00 si no tiene o no reside en la vivienda)
   - conyuge o pareja? En tal caso, dígame su nombre: _____   **NCONY**

   - padre? En tal caso, dígame su nombre: _____   **NPADRE**

   - madre? En tal caso , dígame su nombre: _____   **NMADRE**

5. ¿Cuál es su estado civil legal?   **ECIV1**

soltero	1 ☐
casado	2 ☐
viudo	3 ☐
separado o divorciado	4 ☐
no sabe	0 ☐

Using an electronic questionnaire

2.Optimization of the technological processes-DATA COLLECTION STAGE

- Factors and activities at the ***data collection stage***:

- ◆ 2.1.Using an electronic questionnaire

- ◆ **2.2.Technological Processes**

- ◆ 2.3.Effective field work application. Organization and monitoring of field work

- ◆ 2.4.Weekly download

- ◆ 2.5.Collection and data delivery schedule

2.2. **Technological Processes**

- Customized organization of the technological processes in the collection stage
- The system ensures business continuity at a
  - ◆ database-application-communications level
- Security is ensured in several stages
  - ◆ Login the collection application
  - ◆ Login the interviewers' tablets
  - ◆ Data sending
  - ◆ Intranet
- Process and infraestructure are continously monitored and there is a maintenance support
- Strict policy of backups, both in Delegation servers and tablets and in Central Services

2.Optimization of the technological processes-DATA COLLECTION STAGE

■ Factors and activities at the *data collection stage*:

◆ 2.1.Using an electronic questionnaire

◆ 2.2.Technological Processes

◆ **2.3.Effective field work application. Organization and monitoring of field work**

◆ 2.4.Weekly download

◆ 2.5.Collection and data delivery schedule

2.3.**Field work Application: effective field organization and monitoring**

- The application assists in the collection stage to perform its primary function of collecting and monitoring these data.

- Also, all other associated features such as sample management, resource management, monitoring listings, etc.

- It allows continuous monitoring of field work

- Test and training environment

2.Optimization of the technological processes-DATA COLLECTION STAGE

■ Factors and activities at the *data collection stage*:

◆ 2.1.Using an electronic questionnaire

◆ 2.2.Technological Processes

◆ 2.3.Effective field work application. Organization and monitoring of field work

◆ **2.4.Weekly download**

◆ 2.5.Collection and data delivery schedule

2.4.**Weekly download**

- Once consolidated from all regional offices, information is transferred weekly to the Central Services server.

Phase 0 (LOAD)

- The results of interviews conducted each week in CAPI / CATI are flushed **weekly** to the Central Services servers for centralized processing

- Receiving data

- Checking it and

- Sending reports thereon to the Labour Market unit

- Finally **loading** the survey data to the databases.

2.Optimization of the technological processes-DATA COLLECTION STAGE

- Factors and activities at the *data collection stage*:

- 2.1.Using an electronic questionnaire

- 2.2.Technological Processes

- 2.3.Effective field work application. Organization and monitoring of field work

- 2.4.Weekly download

- **2.5.Collection and data delivery schedule**

21

2.5. Collection and data delivery schedule

- Deadlines for fieldwork and data transfers according to the reference week

- Weekly download, as stablished in the schedule.

- →allows to receive information quickly regarding the reference period of the collection

- →allows the evaluation of the key features of these data received and the potential need for changes at the stage of the data collection

Calendario de la Encuesta de Población Activa. Cuestionario Principal

(Fechas de envios, de cierre, de recepción de listados y de publicación)

Ciclo: 158

Trimestre: 1/2012

Semanas:	1	2	3	4	5	6	7	8	9	10	11	12	13
Envio ordinario mensual o trimestral	24/01/2012	31/01/2012	07/02/2012	14/02/2012	21/02/2012	28/02/2012	06/03/2012	13/03/2012	20/03/2012	27/03/2012	03/04/2012	10/04/2012	16/04/201
Fecha límite de envíos extraordinarios mensuales o trimestrales(*)		Semanas 1 a 4:		23/02/2012		Semanas 5 a 8:		22/03/2012			Semanas 1 a 13:		16/04/201
Fecha de cierre de la EPA mensual o trimestral				24/02/2012				23/03/2012					17/04/2012
Recepción de listados (dep. manual EPA mensual o trimest)		Semanas 1 a 4:		27/02/2012		Semanas 5 a 8:		26/03/2012			Semanas 9 a 13:		18/04/201
Recepción de listados (trat. Automático EPA mensual o trim)		Semanas 1 a 4:		01/03/2012		Semanas 5 a 8:		29/03/2012			Semanas 1 a 13:		23/04/201
Publicación o disponibilidad de las cifras de la EPA mensual o trimestral	EPA de diciembre de 2011:				06/03/2012	EPA de enero de 2012:				03/04/2012	EPA de febrero de 2012: (primer trimestre)		27/04/201

(*) Si la fecha es festiva, el envío se realizará el día anterior

Management of survey processes

Sample Preparation
Workgroup Preparation
Calendar (in weeks)
Workplan (in workgrups-weeks)

Preparing the cycle

Management of the collection

Letters and labels
Workgroup management
Assignation of CAPI work batch
Reserve dwellings
Transfer of dwellings CAPI-CATI
Call control system

Fieldwork

Interviews (CAPI-CATI)
Data Recording
Inspections
Fieldwork notes
Cover and Backcover
Appointments

Monitoring of the collection

Listeners
900 line
Tables and monitoring reports
partial results linstings

closure of cicle and distribution of results

closure of sections and dwellings
cycle rotation
data publication

Centralized Treatment

Host data transfer
Debugging
Imputation

Dump of data from Delegations to Headquarters

Main phases of data process

```
NMOA0014                 INSTITUTO NACIONAL DE ESTADISTICA        DD/MM/YYYY
E572E                          E P A   2 0 0 5
                               PROCESO ORDINARIO
-------------------------------------------------------------------------

                    1.    CARGA CALENDARIO/DICCIONARIO
                    2.    CARGA DATOS EPA          (FASE 0)
                    3.    VALIDACION DATOS EPA     (FASE 1)
                    4.    GENERACION PRE-D.I.A.    (FASE 3)
                    5.    GENERACION D.I.A         (FASE 4)
                    6.    GENERACION POST-D.I.A.   (FASE 5)
                    7.    FASES 7 - 8
                    8.    ACTUALIZACION FICHERO FINAL (FASE 9)
                    9.    CIERRE DE MES/TRIMESTRE
                   10.    MODULO ADHOC PRE-D.I.A
                   11.    MODULO ADHOC POST-D.I.A.



                    Seleccione OPCION (1-11):  __

  DF3   SALID
```

Main phases of data process

- Phase 0.Load. Check register design.

- Phase 1. Automatic validation. Errors in dwelling structure dwelling and control variables.

- Phase 2. Manual Correction of mandatory errors.

- Phase 3-4-5. DIA. Automatic Debugging. Treatment of Responsed and Fault File.

- Phase 6. Validation of auxiliary files.

- Phase 7-8. Creation of the quarterly micro data file. Results.

3.Optimization of the technological processes-DATA PROCESS STAGE

- Factors and activities at the ***data process stage***
  - ◆ 3.1.Availability of auxiliary files
  - ◆ 3.2.Optimization of the procedures used
  - ◆ 3.3.Monthly processes
  - ◆ 3.4.Support systems
  - ◆ 3.5.Availability of resources and coordination
  - ◆ 3.6.Publication Schedule

3.Optimization of the technological processes-DATA PROCESS STAGE

- Factors and activities at the *data process stage*

  - **3.1.Availability of auxiliary files**

  - 3.2.Optimization of the procedures used

  - 3.3.Monthly processes

  - 3.4.Support systems

  - 3.5.Availability of resources and coordination

  - 3.6.Publication Schedule

1. Background

2. Data Collection Stage

3. Data Process Stage

28

3.1. Availability of auxiliary files

- Files that take part in different phases, like the
  - ◆ Geographic Dictionary (with the sections in the sample)
  - ◆ Delivery Schedule,
  - ◆ File of Sections to Repeat and the Population File (to calculate the raising factors)

- These files must be provided on time by the other INE units to carry out the process on time.

3.Optimization of the technological processes-DATA PROCESS STAGE

- Factors and activities at the ***data process stage***
  - ◆ 3.1.Availability of auxiliary files

  - ◆ **3.2.Optimization of the procedures used**

  - ◆ 3.3.Monthly processes

  - ◆ 3.4.Support systems

  - ◆ 3.5.Availability of resources and coordination

  - ◆ 3.6.Publication Schedule

3.2. **Optimization of the procedures used**

- Programming was improved so that processes and database performance were optimized, using an unique key, reorganizing DB2 tables, indexes, etc.

- Process has been adapted to be more dynamic, reducing the phases and therefore the tasks that other units perform

- Inclusion of the DIA

Phase 2 (Manual validation)

```
NMOD0051                    INSTITUTO NACIONAL DE ESTADISTICA              DD/MM/YYYY
                                  E P A   2 0 0 5
                                   IMPUTACIONES
 -----------------------------------------------------------------------------
  NUMORD 01007143
  CICLO 134   PROV 02   SECCION+DC 01 03 4   VIVI 02   NPER 02


  EDAD: 046   BUSCA1: 6   NUEVE2:


  FORB1A:       FORB2A:       FORB3A:       FORB4A:       F5610A:       FOR89A:       FORB7A:
  F1415A:       FOR12A:       FOR16A:       FORB0A:       FORB1B:       FORB2B:       FORB3B:
  FORB4B:       F5610B:       FOR89B:       FORB7B:       F1415B:       FOR12B:


  FOR10C:       FOR56C:       FORB8C:       FORB9C:       FOR14C:       FOR15C:


  DIS132:       EMPANT: 1   RZULT :            OCURN1:




 PF3 - ATRAS
```

- Factors and activities at the ***data process stage***

  - ◆ 3.1.Availability of auxiliary files

  - ◆ 3.2.Optimization of the procedures used

  - ◆ **3.3.Monthly processes**

  - ◆ 3.4.Support systems

  - ◆ 3.5.Availability of resources and coordination

  - ◆ 3.6.Publication Schedule

3.3. **Monthly processes**

- Keypoint to achieve very good timing when calculating quarterly results, because

    - There is an early detection of potential problems or inconsistencies in data input.

    - Part of the quarterly registers are already depurated when quarterly process begins

    - Programs are checked out every month and problems detected in the 1-8 weeks are already solved.

- Other advantage: Providing a file for the Labour Market unit to be used in monthly estimations

- Factors and activities at the ***data process stage***
  - ◆ 3.1.Availability of auxiliary files

  - ◆ 3.2.Optimization of the procedures used

  - ◆ 3.3.Monthly processes

  - ◆ **3.4.Support systems**

  - ◆ 3.5.Availability of resources and coordination

  - ◆ 3.6.Publication Schedule

3.4. **Support systems**

- In case communications or computer systems in the INE, fail.

- Ensure business continuity by operating against a mainframe in another environment

- Processes to be undertaken in Central Services are replicated.

- Factors and activities at the *data process stage*

  ◆ 3.1.Availability of auxiliary files

  ◆ 3.2.Optimization of the procedures used

  ◆ 3.3.Monthly processes

  ◆ 3.4.Support systems

  ◆ **3.5.Availability of resources and coordination**

  ◆ 3.6.Publication Schedule

3.5. **Availability of resources and coordination**

- Human resources are available for application maintenance and update, either to develop new functionalities that have to be included in the process, or to troubleshoot process issues.

- Coordination in the different IT units involved in the processes

3.Optimization of the technological processes-DATA PROCESS STAGE

- Factors and activities at the ***data process stage***
  - ◆ 3.1.Availability of auxiliary files

  - ◆ 3.2.Optimization of the procedures used

  - ◆ 3.3.Monthly processes

  - ◆ 3.4.Support systems

  - ◆ 3.5.Availability of resources and coordination

  - ◆ **3.6.Publication Schedule**

1. Background

2. Data Collection Stage

3. Data Process Stage

3.6. **Publication Schedule**

- Shows the dissemination dates for the quarterly results.

- These results are published **two weeks after the closing date of the quarter**.

- This deadline requires the data process optimization.

1. Background

2. Data Collection Stage

3. Data Process Stage

Results

- From the final quarterly microdata file, output of the 7 Phase,

  - the results **tables** are obtained to publish,

  - the files are prepared for **distribution** by various means (PC-AXIS, TEMPUS database, Web) and

  - **anonymised files** are generated for users of regular requests.

  - There also are obtained internal **review boards**.

Data Dissemination

❑ Quarter 1: 4th week of April.
 Quarter 2: 4th week of July.
 Quarter 3: 4th week of October.
 Quarter 4: 4th week of January the following year.

❑ http://www.ine.es/inebmenu/indice.htm

Anexes

Phase 0 (LOAD)
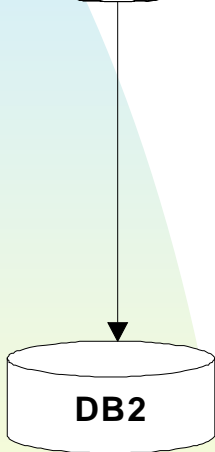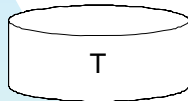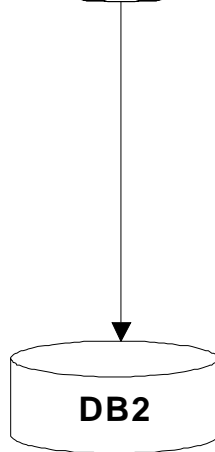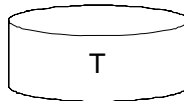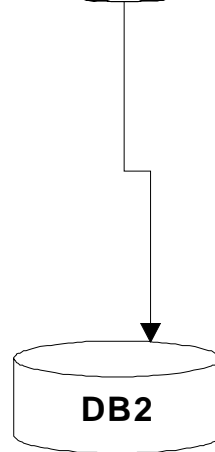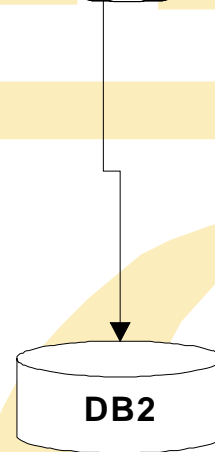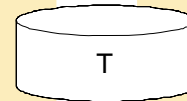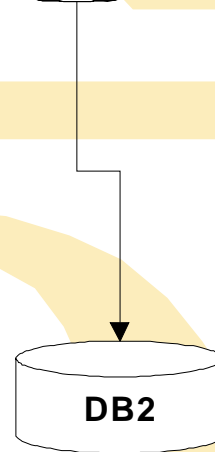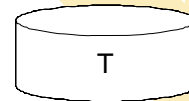
Phase 6 (POBYSECC)

- Validation of the **file of sections to be repeated**.
- Validation of the **population files** of the month for publication.
- Calculation of **repetition factors** (FACREP) and **elevation** factors (FACELE).
- If all identifications of sections to be repeated are in the **dictionary** and all have a **repetition factor** greater than or equal to 2 the file is **cataloged**, otherwise they'd be considered critical errors

Data delivery schedule

■ **Calendar**:

- ◆ There is a schedule for each quarter of closure and dissemination dates

- ◆ In between, should be the data processing.

- ◆ The **closing date of a month** is the reference date for the monthly process

- ◆ What If after that date and before the close of the quarter we receive new data for that month?

- ◆ The **closing date of a quarter**, indicates the start the process of the **quarterly treatment**.

Main phases of data process

- Although EPA is a quarterly survey, the centralized data processing consists of weekly, monthly and quarterly phases.

- Why Monthly phases?(No dissemination)

- The results for a quarter represent the central month (second) for that quarter

- Also, for the estimation of the monthly processes results

- **Exploitation** (tabulation of results, preparation of the database load, anonymisation of files, requests, etc ....) of the final quarterly microdata file is made in the SGTIC and is **part of the quarterly processes**.

48

Geographic Dictionary

DGEPAT212.TXT - Bloc de notas

Archivo   Edición   Formato   Ver   Ayuda

```
1601ALAVA              0029Amurrio               01001 0306477159      1101022  60
1601ALAVA              0040Artziniega            01001 0304217159      2501023  60
1601ALAVA              0339Lapuebla de Labarca    01001 0311327159      2601043  60
1601ALAVA              0360Laudio/Llodio         01013 0202406159      1101022  10
1601ALAVA              0360Laudio/Llodio         01004 0207346159      2601022  60
1601ALAVA              0360Laudio/Llodio         01008 0209586159      2201022  10
```

Technological Environment

- IBM Mainframe 2064/102 with 449 MIPS and 8GB of RAM

- Operating System Z / OS v1.8

- Database Manager: DB2 V 8.1

- Programming Languages: NATURAL, PL1, SAS, TPL (language of tabulation and aggregation of data)

- Support Information: flat files, VSAM, and DB2 database

- Batch execution Languages   : JCL

```
┌─────────────────┐
│     PreDIA      │
└─────────────────┘
         ↓
┌─────────────────┐
│       DIA       │
└─────────────────┘
         ↓
┌─────────────────┐
│     PostDIA     │
└─────────────────┘
         ↓
┌─────────────────┐
│ Other Processes │
└─────────────────┘
         ↓
      Final
      microdata file
```
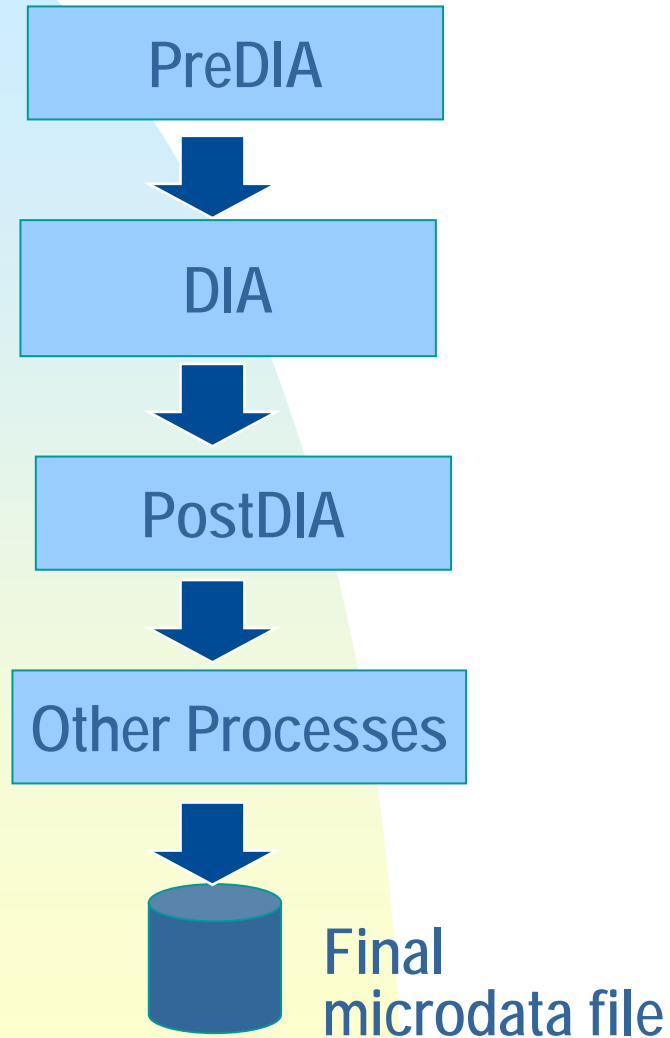
Phase 4 (D.I.A.) Description

- ❑ Application developed by the INE for the treatment of qualitative variables.

- ❑ Based on the Fellegi and Holt methodology with modifications to handle systematic errors.

- ❑ Allows only detection or detection and imputation

- ❑ Consists of two independent **subsystems**:

  - ❑ The **random imputation**. It is based on *edits*.

  - ❑ The **deterministic** imputation. It is based on *rids*.

- ❑ By the first subsystem there will be made imputations according to the following principles:

  - ❑ It has to respect the original distributions of the variables (ie it is assumed that errors are random).

  - ❑ Be maintained the maximum of original information (principle of minimal change).

Phase 4 (D.I.A.) Description

❑ **Advantages** of the DIA:

   ❑ Ease of use: just set the inconsistencies.

   ❑ Flexible: easily modify the rules.

   ❑ Provides plenty of information: imputations, distributions, error rates, etc..

❑ **Disadvantages** of the DIA:

   ❑ Does not support rules of inconsistency between variables of different records

   ❑ Time consuming process and computer memory

1.
Background

2. Data
Collection
Stage

**3. Data
Process
Stage**

53

Specifications

DIA

DIABEGIN

DIACRFIC

DIA1CD

…

DIA8IMPU

DIATRATA

Data

Correct Data   Imputed data   Lists

Automatic Debug: Preparation of DIA

Before the Quarterly process

1.- Specifications Files are introduced in the system:

Debugging variables
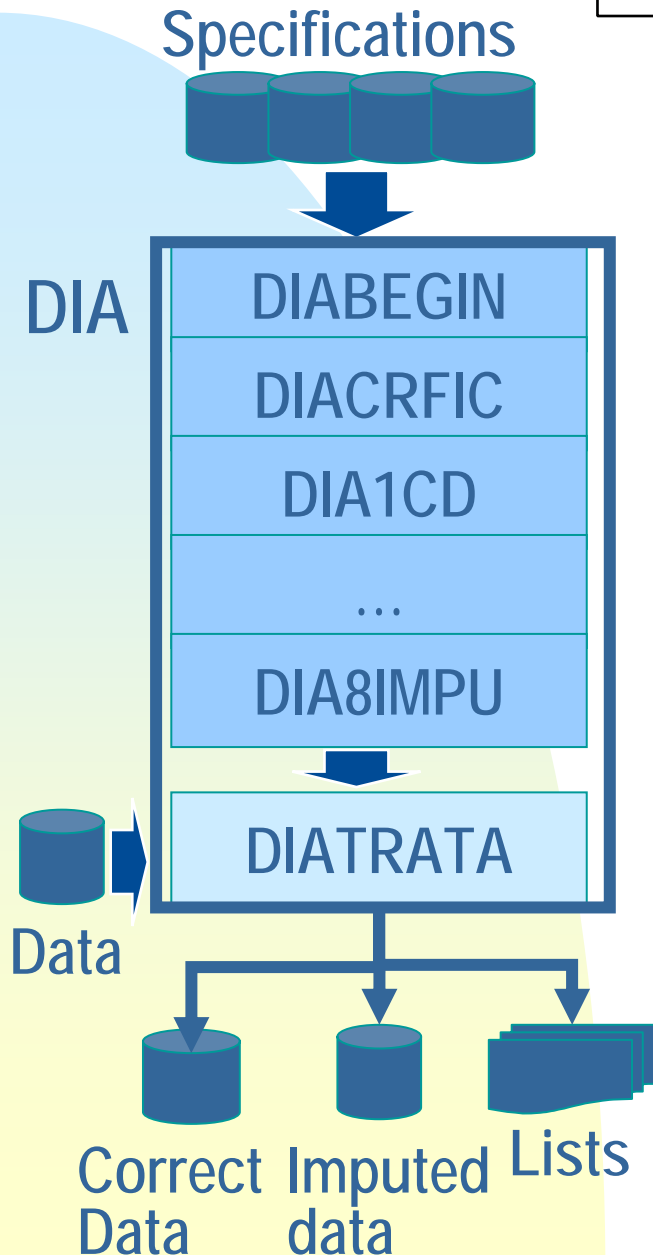
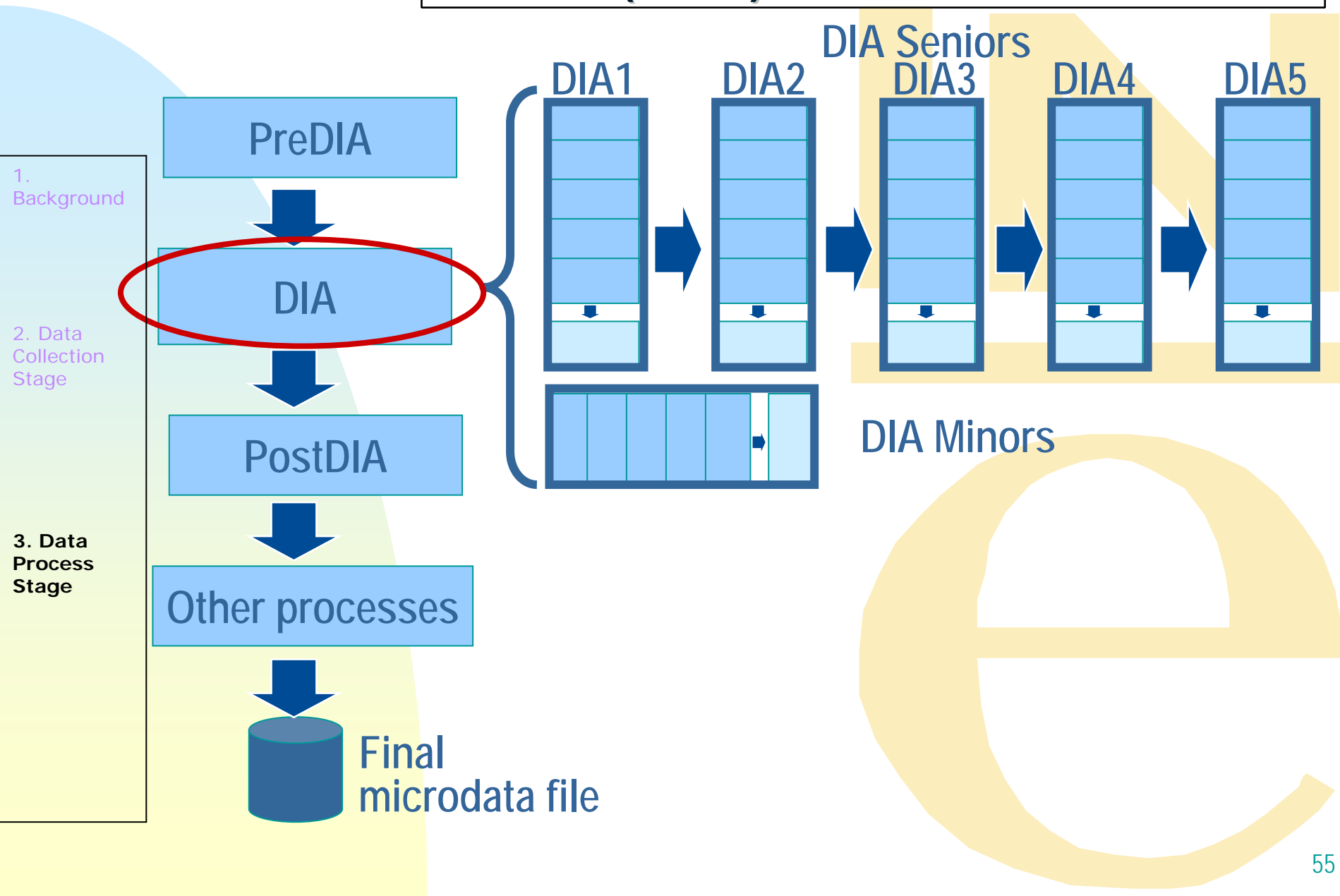Position and length of the variables

rids

edits

fixed fields

Valid values   ...

2.- All phases of the DIA are run except the last one, and the internal files are generated, so that in the quarterly process will enter the process in the DIATRATA with the data to be debugged

Phase 4 (D.I.A.)

DIA Seniors

DIA1   DIA2   DIA3   DIA4   DIA5

PreDIA

1. Background

DIA

2. Data Collection Stage

PostDIA

3. Data Process Stage

Other processes

Final microdata file

DIA Minors

Population 16 years and over by sex and relationship to the economic activity

Resultados nacionales

	Trimestre actual	Variación sobre el trimestre anterior		Variación sobre igual trimestre del año anterior	
		Diferencia	Porcentaje	Diferencia	Porcentaje

1. Población de 16 años y más por sexo y relación con la actividad económica

AMBOS SEXOS					
Población de 16 años y más	38,508.2	20.4	0.05	-4.3	-0.01
Activos	23,081.2	-53.4	-0.23	-23.6	-0.10
- Ocupados	17,807.5	-348.7	-1.92	-600.6	-3.26
- Parados	5,273.6	295.3	5.93	577.0	12.29
Inactivos	15,427.0	73.8	0.48	19.4	0.13
Tasa de actividad	59.94	-0.17	-	-0.05	-
Tasa de paro	22.85	1.33	-	2.52	-
Población de 16 a 64 años	30,658.5	-20.8	-0.07	-153.8	-0.50
Tasa de actividad (16-64)	74.75	-0.18	-	0.27	-
Tasa de paro (16-64)	23.00	1.35	-	2.54	-
Tasa de empleo (16-64)	57.56	-1.15	-	-1.68	-

56

Ocupados y parados por sexo

EPA   Cuarto Trimestre 2011

Ocupados

Parados