QÜESTIIÓ, Vol 19, 1, 2, 3 pp. 261-320, 1995

INTRODUCCIÓ ALS SISTEMES DE METAINFORMACIÓ ESTADÍSTICA

ISIDRE CANALS CABIRÓ*

D'una banda, es defineix el concepte i la tipologia de la metainformació estadística i es descriuen els problemes plantejats als instituts d'estadística per a la seva organització en sistemes integrats, de cara a la satisfacció de les necessitats dels usuaris d'informació estadística, tant externs com interns.

D'altra banda, es passa revista als diferents projectes i experiències d'altres instituts agrupats segons els enfocaments tècnics possibles, incloent-hi els plantejaments i recomanacions dels grups de treball específics d'institucions internacionals.

Finalment, es treuen algunes conclusions i criteris pràctics tant per definir una estratègia com per dissenyar un sistema global de metainformació estadística.

Aquest article s'ha de considerar introductori a la problemàtica i complementari de l'article de Sundgren, també publicat en aquest número de ***Qüestiió****.*

Introduction to Statistical Metainformation Systems.

Key words: Metainformació estadística, Metadades, Sistema de metainformació estadística, Sistema d'informació estadística.

* Isidre Canals Cabiró. Institut d'Estadística de Catalunya. Dept. d'Economia i Finances. Generalitat de Catalunya. Via Laietana, 58. 08003 Barcelona.

Les opinions expressades en aquest article no reflecteixen necessàriament les de l'Institut.

- Article rebut el maig de 1995.
- Acceptat el setembre de 1995.

1. INTRODUCCIÓ A LA METAINFORMACIÓ ESTADÍSTICA

La manera més simple de definir la metainformació estadística és la d'identificar-la amb la *informació sobre la informació estadística.*

Resulta també la més interessant des de la perspectiva d'un institut d'estadística. En tot cas, més general que el significat corrent del terme metadades, quan es refereix només a la informació relativa a dades numèriques concretes (tant si es tracta de *microdades*, és a dir, dades individualitzades, com de *macrodades*, és a dir, dades agregades en forma de taules)[1].

En efecte, per a un institut d'estadística, el significat del terme "metainformació estadística" inclou també tot el que es refereix a informació de l'estadística en general (concretat, però, en la pràctica, al país de referència i a l'àmbit de les competències de l'institut de què es tracti). Plantejada així la qüestió, resoldre el problema de l'organització de la metainformació estadística resulta vital, no solament pel que fa a l'organització de la informació estadística al servei de les necessitats internes d'un institut d'estadística, sinó també pel que fa a l'articulació d'unes relacions eficients entre l'institut i la societat a la que serveix.

Així, no semblarà estrany que quan Gillman, del *U.S. Bureau of the Census*, presentava la seva visió del problema de la metainformació, l'encapçalava amb aquestes paraules, amb les que s'havia expressat oficialment quin hauria de ser l'objectiu global a perseguir a l'era de la revolució de les tecnologies de la informació i la comunicació, pel que fa a l'estadística: "*un sistema de recollida, producció i difusió de dades electròniques sense discontinuïtats entre les seves fases i íntimament entrelligat amb la societat a la que mesura*". (Gillman 1994).

Des d'una perspectiva més operativa, dos són els punts de vista amb els que, al si d'un institut d'estadística, pot ser plantejada la metainformació estadística, i que corresponen *grosso modo*: el primer, al punt de vista dels usuaris (i de l'*ús*) de la informació estadística, reprès pels responsables i tècnics (temàtics, informàtics i especialistes d'informació) encarregats de la difusió d'estadístiques, i el segon, al punt de vista de la *producció*, encarnat pels responsables i tècnics (estadístics, temàtics i informàtics) al càrrec de la producció d'informació estadística.

Aquests dos punts de vista indueixen no solament plantejaments diferents respecte de la problemàtica de la metainformació estadística, sinó que responen a necessitats

[1]Cal advertir, però, que la tasca de recerca terminològica i d'estandardització empresa pel grup de treball METIS (de la Conferència d'Estadístics Europeus, dedicat als problemes de la metainformació estadística) distingeix entre "*statistical metainformation*" (= *An information on statistical metadata or statistical metainformation*) i "*metadata*" (= *Data representing statistical metainformation*). Vg [Prazenka(1994)]. Vg també l'apartat 2.4.2).

diferents, segons les tasques i coneixements dels usuaris (interns i externs), que requeriran informacions diferenciades sobre la mateixa qüestió. Aquests plantejaments, però, estan destinats a convergir en el futur, en la mesura que responen a funcions complementàries en tot institut d'estadística i, d'altra banda, l'organització de la informació haurà de tendir a presentar-se als usuaris de tota mena amb un únic accès i una conceptualització global consistent.

La convergència provindrà també de l'ampliació del terme "usuaris" per comprendre també (a més dels usuaris finals) tota mena d'usuaris interns. D'aquesta manera, la metainformació estadística ja no serà només aquella informació complementària, necessària per a un *ús final* correcte de les dades, sinó també (i, des del punt de vista de la gestió d'un institut d'estadística, sobretot) aquella informació necessària per a la *producció* eficient de les dades. Podríem parlar, així, d'una metainformació estadística final i d'una metainformació estadística instrumental.

Afegim que, al llarg d'aquest article, el terme "usuaris" manté una certa ambigüetat, donat que, segons el context d'on provingui la informació descrita, el seu significat pot anar del més restrictiu al més ampli. Confiem que el bon sentit del lector sàpiga discernir en cada part el significat més adequat.

1.1. La metainformació estadística des del punt de vista de l'ús de la informació estadística

Una primera consideració ens porta als orígens de la metainformació estadística, als que fa referència el terme anglès *metadata*, utilitzat per primera vegada fa una vintena d'anys per Sundgren[2] (Institut Suec d'estadística) i que identifica aquella part de la metainformació més propera a les dades pròpiament dites.

En efecte, la necessitat de les metadades hauria nascut per la preocupació (que s'ha manifestat tostemps, amb més o menys força) de que les dades estadístiques vagin sempre acompanyades de les informacions necessàries per a interpretar-les correctament. Recordem que les dades estadístiques, fins i tot les més simples, són sempre estimacions, i que només es poden determinar els límits entre els que es troba el valor real o veritable d'una variable. I, així, les notes que apareixen al peu d'una taula, o annexes a ella, constitueixen la manifestació més elemental de la metainformació estadística: definicions, característiques de la mostra d'una enquesta, diferències en la data o període cobert, mètode emprat per a la recollida o l'estimació de valors, incidències en el treball de camp, etc.[3]

[2]Bo Sundgren és reconegut com un pioner en l'estudi i plantejament conceptual de la meta-informació i els sistemes de metainformació estadística.

[3]Vg. Silver (1993), on fa un plantejament sistemàtic de les notes al peu de taules estadístiques, i on es

Del que es tracta, en definitiva, és de proporcionar a l'usuari, en una publicació estadística, a l'ensems de les dades el seu context, condicionant decisiu de la fidelitat amb la que reflecteixen el fenòmen de la realitat que pretenen medir. Només així els usuaris de l'estadística podran estar en disposició d'utilitzar- les correctament.

Aquesta materialització de la metainformació estadística en notes acompanyants de les taules estadístiques en publicacions impreses continúa essent molt important; en l'actualitat, però, cal tenir en compte dos nous mitjans de difusió de la informació estadística: l'*edició electrònica* i l'*accés interactiu a bases de dades* públiques, que hem de considerar també des de la perspectiva de la metainformació.

Comencem per constatar que ha començat amb força creixent i irreversible l'edició d'informació estadística en forma digitalitzada i en suports magnètics (cintes i disquets) i òptics (CD-ROM). I aquest fet obliga a plantejar la problemàtica de les metadades amb característiques noves. D'una banda, pel problema dels formats sota els quals hauran de ser distribuïdes la informació estadística i la metainformació associada (en el que no entrem ara, però és decisiu a llarg termini). De l'altra (sobretot), per les enormes possibilitats obertes en l'organització conjunta d'ambdues menes d'informacions en sistemes interactius, provistos de les funcionalitats vehiculades per les interfícies gràfiques d'usuari[4].

L'altre mitjà de difusió de la informació estadística consisteix a oferir al públic l'accés interactiu a bases de dades[5]. Aquests sistemes, de les seves primeres manifestacions als anys 70 ençà, han passat per vàries generacions, pel que fa a les seves funcionalitats. Avui dia, es troben en plena crisi de transformació, coexistint inharmònicament diverses formes, en conflicte i al mateix temps convergents: bases de dades documentals de recuperació d'informació clàssiques (*retrieval systems*), bases de dades relacionals (ambdues eventualment dotades d'interfícies gràfiques d'usuari), servidors *World-Wide-Web (WWW)* al si de la xarxa Internet (que utilitzen tècniques hipertext simples), etc.

Així, doncs, ens cal preguntar-nos com s'ha d'integrar la metainformació en aquests mitjans digitalitzats. La resposta haurà de tenir en compte, d'una banda, les més grans potencialitats dels sistemes informatitzats, però també, de l'altra, el fet que els usuaris de la informació estadística també han evolucionat i les seves neces-

proposa una taxonomia específica en 22 menes de notes, codificades i agrupades en 5 grups principals.

[4]Les interfícies gràfiques d'usuari (*GUI = Graphical User Interfaces*) corresponen a tot un paradigma informàtic, altrament identificat amb les lletres *MIWM*, que corresponen als seus quatre components essencials: *Menu-Icon-Window-Mouse*), també anomenats *sistemes de manipulació directa* (Shneiderman).

[5]El *Consorci d'Informació i Documentació de Catalunya*, precedent històric i base de l'actual *Institut d'Estadística de Catalunya*, fou pioner a Espanya, tant en l'accés a les bases de dades en línia científiques i tècniques del primer servidor europeu (*European Space Agency*), com en la producció i posada en servei públic de la primera base de dades estadística (ESPAN) sobre les fonts estadístiques espanyoles, i, en general, en l'aplicació de la informàtica als problemes documentals.

sitats són més diversificades i complexes. De manera que no n'hi haurà prou amb preveure simples notes acompanyant les dades. D'una banda, les peces de metainformació hauran d'incloure no solament textos llargs (legislació, p. ex.) sino també documents composats (projectes tècnics, incloent gràfiques, taules, cronogrames, etc.) i, fins i tot, bases de dades estructurades (com en el cas de les Classificacions multinivells). Tindrem així un conjunt voluminós i heterogeni de peces d'informació, l'organització de les interrelacions entre les quals no és gens simple. D'afegitó, els usuaris ja no s'acontentaran amb la simple consulta de la informació; voldran (i és raonable que ho esperin així) importar-la al seu ordinador i integrar-la al seu propi entorn de treball de la manera més simple possible.

El resultat final és el de que un institut d'estadística ha de preveure l'emmagatzemament (o elaboració ad hoc) d'una munió de peces de metainformació heterogènies i organitzar-les en un veritable sistema de metainformació estadística, provist de les funcionalitats adequades als fins a què es destina.

La natura tècnica d'aquests sistemes d'informació pot correspondre a la dels sistemes clàssics de recuperació d'informació, però també pot adoptar avantatjosament la dels sistemes hipertext.

Efectivament, els sistemes hipertext, altrament dits navegacionals per la seva capacitat de permetre l'usuari *navegar* al seu gust per entre les peces d'informació (*nodes*), atorguen una gran llibertat al dissenyador a l'hora d'interrelacionar-les mitjançant *lligams* (*links*). I aquesta és una característica especialment adequada quan es tracta d'interrelacionar la informació amb la metainformació estadística. Justament, però, per aquella llibertat, és molt important estudiar amb cura els principis a què haurien de respondre l'estructura dels lligams i els mapes conceptuals constitutius del *browser*, com s'anomena l'instrument bàsic de navegació dels usuaris[6].

En resum, la perspectiva dels usuaris de la informació estadística ens porta a organitzar sistemes de recuperació i difusió com a eines molt adequades per a satisfer les necessitats modernes d'informació sobre la informació estadística. És més, *la perspectiva de l'ús pot ser utilitzada fins i tot a l'hora d'organitzar el propi sistema d'informació estadística d'un institut.* Si així es fa, aquest és anomenat per Sundgren (1993) *sistema d'informació estadística orientat a l'usuari* (*user-oriented*) o bé *induït per l'usuari* (*user-driven*).

Pel que fa a la metainformació estadística, aquella generalització i diversitat de mitjans a disposició del públic planteja la possibilitat i l'exigència addicional, per als dissenyadors dels sistemes, de considerar les necessitats específiques de grups

[6]Vg Canals (1990), on es tracta aquest problema específic, amb caràcter general, per a qualsevol mena d'informació en els sistemes hipertext.

d'usuaris diferenciats, i en especial la del grup d'usuaris sense coneixements tècnics amplis; la qual cosa es tradueix en la incorporació de peces de metainformació estadística de tipus explicatiu i pedagògic, a més de les de caire fonamentalment tècnic.

Finalment, no cal oblidar la necessitat més general i evident dels usuaris potencials d'informació estadística: la de conèixer les fonts estadístiques disponibles sobre un tema determinat. Així, la descripció de les fonts d'un país hauria de constituir una peça bàsica en l'articulació del sistema general de metainformació estadística corresponent al seu Sistema estadístic.

1.2. La metainformació estadística des del punt de vista de la producció d'informació estadística

Si ara ens situem en el lloc dels responsables i tècnics de tota mena involucrats en la *producció* d'informació estadística que és, evidentment, la funció bàsica de tot institut d'estadística, la metainformació estadística que ens interessarà serà tota aquella informació addicional necessària per produir una informació estadística en particular, o un conjunt d'informacions estadístiques en general (a banda, naturalment, de la informació numèrica mateixa). Així, totes aquelles informacions relatives a totes i cadascuna de les fases de la cadena de producció de la informació estadística (disseny, obtenció de les dades primàries, constitució dels arxius de base, estimació, explotació[7]) constituïran peces de metainformació pertinents, independentment de la seva forma i suport del sistema d'informació en el que puguin estar integrades.

Així, per començar, tota la *documentació tècnica* relativa a les operacions estadístiques constituirà una part substancial de la metainformació estadística, com també la informació legislativa i l'administrativa associada o relacionada. Un altre conjunt típic és el de la *informació instrumental*, entre la que podem distingir la relacionada amb el tractament informàtic (documentació de sistemes, programes, codis, etc.) i la normativa (classificacions, codis territorials, etc.).

Amb aquesta simple enumeració es comprèn a la vegada la complexitat i la trascendència per a un institut d'estadística d'organitzar eficientment aquesta munió de peces d'informació heterogènia, de manera que el seu accés, consulta i eventual reutilització siguin no solament factibles sinó facilitats al màxim.

Un primer desenvolupament informàtic (*Data dictionaries*) ha vingut a oferir-se com a una solució tecnològica específica, al menys per a les necessitats d'informació instrumental comuna a un conjunt de bases de dades al si d'un mateix sistema. Que puguin constituir el fonament tecnològic d'un sistema complet de metainformació es-

[7] Als que caldria afegir la *difusió* i, fins i tot, l'eficàcia social en l'*ús* de la informació estadística.

tadística, però, està per veure, donada l'heterogeneïtat de les peces d'informació que aquest ha de tractar, la diversitat de les funcionalitats reclamades i la dispersió dels nuclis d'arxiu existents, que necessàriament s'han d'integrar en aquell sistema, i que ultrapassen clarament els objectius dels diccionaris de dades.

Des de la perspectiva del seu ús, és evident que les diferents menes de metainformació de què ara parlem estan concebudes per estar bàsicament al servei del personal relacionat amb la producció en un institut d'estadística. Per això, els sistemes que l'organitzen són anomenats per Sundgren *sistemes de metainformació estadística orientats o induïts per la producció* (*production-driven*), marcant així una diferència important amb l'orientació dels sistemes de què parlàvem a l'apartat anterior, al servei prioritari dels usuaris (externs sobretot) de la informació estadística.

Ara bé, dit això, convé afegir que les noves possibilitats de la telemàtica, entre d'altres coses, obren perspectives inèdites, que tenen com a conseqüència que aquesta metainformació estadística, induïda per les tasques de la producció, s'apropi i arribi a confondre's (al menys en part) amb la reclamada per les necessitats dels usuaris finals d'informació estadística.

En efecte, en primer lloc, si considerem, com és lògic i cada vegada més estès, que la feina d'un institut d'estadística no s'exhaureix amb l'explotació, sinó que inclou també la difusió (i sobretot la difusió activa, més enllà d'unes quantes publicacions impreses), tot el que hem dit dels sistemes *induïts pels usuaris* a l'apartat 1.1 pot ser reprès per ser integrat ara amb la perspectiva interna d'un institut d'estadística.

En segon lloc, avui dia, entre els usuaris externs, abans poc versats en els tecnicismes estadístics, hi ha cada vegada més usuaris amb els coneixements tècnics i la capacitat i els instruments necessaris no solament per interpretar correctament les dades, i eventualment reutilitzar mètodes i peces d'informació instrumentals, sinó també per explotar pel seu compte arxius de microdades (en la mesura que puguin ser difosos en forma no contradictòria amb les restriccions del secret estadístic).

En tercer lloc, les exigències de coordinació (i fins i tot de cooperació) entre instituts porten a permetre l'accés dels tècnics d'uns instituts a la informació tècnica dels altres. I això, tant al si del sistema estadístic d'un país[8], com a nivell de la coordinació europea o internacional entre instituts d'estadística[9].

---

[8]Un exemple n'és el *Sistema Estadístic de Catalunya*, estructurat entorn del *Pla Estadístic de Catalunya* (*PEC*), com a eina legislativa bàsica de planificació quadriennal, i al si del qual l'*Institut d'Estadística de Catalunya* exerceix un paper fonamental de coordinació tècnica entre els diferents organismes que hi participen, essent-ne garantia de la seva correcta execució.

[9]Un exemple n'és la coordinació a nivell de la Unió Europea, que ha de rebre en els propers anys un impuls decisiu amb la implementació del projecte *DSIS* (*Distributed Statistical Information System*) d'EUROSTAT (1992).

1.3. Perspectiva històrica

Arribats aquí, pot ser instructiu seguir l'evolució històrica que ha seguit la metainformació estadística.

Reformulant les fases definides per Nordbotten (1993), que descriuen el desenvolupament històric de mètodes de tractament de la metainformació estadística, podem definir sis fases relativament diferenciades (*Vg. Figura 1*).

En primer lloc, recordem que s'ha reconegut sempre, des dels inicis de la producció d'estadística oficial, la necessitat d'acompanyar les dades (extretes, per exemple, dels censos primitius) amb textos descrivint l'abast i els procediments i mètodes utilitzats per a la recollida de les dades. Les metadades més comunes eren, i probablement encara són avui dia, les notes al peu de les taules estadístiques.

En una segona fase, cap als anys 50, amb les primeres aplicacions de sistemes electro-mecànics de tractament de dades, certs detalls metodològics (si bé molt codificats i críptics) podien ser inclosos a col·leccions de targes perforades, les quals podien així anomenar-se arxius autoexplicatius (*self-descripting files*), encara que només amb un abús del llenguatge podríem avui dia admetre aquesta terminologia.

Només en una tercera fase, als anys 60, amb la nova generació d'ordinadors i la generalització de la cinta magnètica, és possible parlar d'arxius de dades sistemàtics (*data archives*), que podien ser objecte d'intercanvi. Aquests arxius incloïen codis i notes de metainformació, si bé tan succints, que obligaven a acompanyar-los d'una documentació explicativa a part, en paper.

Els reptes que representaven les noves possibilitats informàtiques (tot i restringides per la gairebé exclusiva consideració de camps de longitud fixa) varen fer l'objecte de grans debats al si de la comunitat estadística, encapçalats pels treballs metodològics de la Conferència d'Estadístics Europeus, a través del "*Working Group for EDP (Electronic Data Processing)*".

Ara bé, com fa notar Sundgren amb força, cal recordar que la informatització de la gestió de les operacions estadístiques va implicar en la pràctica, i des dels seus inicis, la desafortunada conseqüència de la desintegració (separació) de la relació natural que havia existit entre dades i metadades als sistemes manuals. En efecte, històricament, les preguntes, instruccions i respostes d'un qüestionari no es separaven en cap dels processos de la cadena manual; la mecanització, en canvi, que es limitava als processos de recompte (numèrics), només recollia les dades nues[10].

[10]Aquesta conseqüència, negativa, de la gairebé exclusiva dedicació al càlcul numèric de la informàtica tradicional, es retroba en molts camps (p. ex. en la informàtica documental), malgrat algunes veus aïllades, com la de Ted Nelson, pioner de l'hipertext, que als anys 60 propugnava les *literary machines* enfront dels *computers*.

abans de 1950	**Descripcions textuals.**
1950	**Autodescripcions per interpretació de targes perforades.**
1960	**Arxius de dades estadístiques (*data archives*). Només codis.**
1970	**Sistemes d'informació estadística. Concepte de metadades.**
1980	**Sistemes de metainformació estadística.**
1990	**Experiències diverses (producció *vs* ús). Consciència de complexitat i globalitat.**

Figura 1
Fases històriques de la metainformació estadística.

Només va ser en una quarta fase, als anys 70, quan s'obre pas el concepte de sistemes d'informació estadística, superant el d'arxius individuals, i la metainformació estadística comença a ser estudiada com un problema específic pels instituts d'estadística, sobretot per Sundgren (1973) en els seus treballs seminals.

En els anys 80, en una cinquena fase, comença a estar clar que la problemàtica de la metainformació estadística era més complexa del que havia semblat, i calia desenvolupar noves eines conceptuals i tècniques. D'altra banda, en una època de turbulència tecnològica, en la que coexistien grans sistemes centralitzats en temps real i xarxes locals de PCs, els esforços i les inversions es concentren en els instituts en el desenvolupament de sistemes d'informació estadística, deixant de banda la metainformació estadística. Tot i així, el debat sobre les metadades no s'atura i, fins i tot, comencen a desenvolupar-se alguns sistemes de metainformació estadística, si bé parcials.

Finalment, en una sisena fase, en la que ens trobem, en ple canvi de paradigma informàtic, sembla, d'una banda, que la tecnologia necessària per a la integració i la navegació entre informacions i sistemes heterogenis comença a estar disponible (orientació a l'objecte, hipertext/hipermèdia, interfícies gràfiques d'usuari), però les aplicacions divergeixen entre sistemes orientats a la producció o als usuaris, tant en els conceptes com en les plataformes utilitzades (Sadreddini 1993).

En contraposició a la *desintegració* entre les dades i les metadades patida fins ara com a resultat de la informatització, Sundgren subratlla que "una característica essencial de la gestió moderna de les meta dades és que està[11] *reintegrada* amb la gestió de les dades objecte".

D'altra banda, s'ha reafirmat la consciència de la manca d'un model conceptual adequat per al tractament i la integració de la metainformació estadística. I, sobretot, s'admet l'absoluta necessitat que qualsevol pas endavant haurà de ser fet en perfecta coordinació internacional. La globalització de l'economia no permet fer-ho altrament. I això implica haver resolt molts problemes d'homogeneïtzació de formats, conceptes, sistemes i xarxes, més enllà dels purament lingüístics.

En aquest sentit, si al febrer de 1993 se celebrava el 1r *Statistical MetaInformation Systems Workshop*, organitzat per EUROSTAT[12], i al 1994 s'aprovaven les grans línies del *IDSIS (Distributed Statistical Information Systems)*, el 1995 ha vist el desplegament al gran públic de l'estratègia de les "*superautopistes de la informació*", que plantegen noves i fortes exigències als vells i no resolts problemes de la metainformació, però també ofereixen enormes potencialitats.

1.4. Tipologia de la metainformació estadística segons la seva natura

1.4.1. Diversitat de la metainformació estadística

Si centrem ara la nostra atenció en les diferents menes de metainformació estadística que caldria organitzar hipotèticament en un sistema, ens trobem que una característica important del conjunt que podríem llistar és la seva diversitat, característica que constitueix un dels problemes bàsics a resoldre.

Diversitat tant en el que fa a la natura de la pròpia informació com en el que fa als formats en els que es presenta, com a les característiques dels suports que la vehiculen.

Respecte de la natura, i en una primera anàlisi, un criteri decisiu és el que separa, d'una banda, la metainformació que es refereix a *microdades* (bé a les dades mateixes, emmagatzemades en un arxiu individualitzat, bé a qualsevol de les fases de l'operació estadística que les ha creat) i, de l'altra, la que es refereix a *macrodades*, és a dir, les taules resultat de l'explotació de les microdades.

[11] No sempre, però seria desitjable que així fos, afegim nosaltres.

[12] La lectura de les actes d'aquest *Workshop* constitueix una magnífica oportunitat d'apropar-se a la problemàtica de la metainformació estadística (EUROSTAT (1993).

La metainformació sobre microdades[13] serà, més aviat, d'interès per als tècnics d'un institut d'estadística, mentre la referent a les macrodades anirá més aviat destinada als usuaris de les dades, normalment externs a l'institut d'estadística.

1.4.2. Metainformació sobre les microdades

Ara bé, dintre del primer grup, podem identificar menes d'informació molt diferents segons la seva natura, des de codis i detalls técnics molt concrets utilitzats pels informàtics per al tractament dels arxius en el transcurs d'una operació estadística, fins a la documentació tècnica relativa al disseny d'un qüestionari o d'una operació de camp (recollida), com també al detall de les incidències ocorregudes durant la seva implementació, passant per les consideracions (des dels punts de vista temàtic o bé estadístic) relatives a les variables i característiques de la població objecte de quantificació o mesura.

Dintre encara d'aquest primer grup, és a dir, la metainformació sobre les microdades, trobem les menes de metainformació que es refereixen o s'utilitzen en un marc més ampli al d'una operació estadística concreta (o d'una sèrie d'elles): ens referim, per exemple, al conjunt de programes informàtics, tant aplicatius com de base (llenguatges o bases de dades), i també a una mena de metainformació extraordinàriament important per a l'estadística: les Classificacions i Nomenclatures.

Aquestes darreres, per la seva pròpia natura, pels contextos internacional i multilingüístic de la seva creació i la complexitat del seu tractament multinivells jeràrquics, plantegen problemes específics, que justifiquen per a molts instituts d'estadística el seu tractament autònom (i, fins i tot, prioritari dins la metainformació estadística).

Altres menes de metainformació general cauen també dintre d'aquest primer grup, com poden ser la legislació i la reglamentació administrativa, quan es refereixen a una o vàries operacions estadístiques, o bé a tot el sistema estadístic o a alguna de les seves parts.

1.4.3. Metainformació sobre les macrodades

En el que fa al segon grup, és a dir, la metainformació sobre les macrodades, ha estat objecte d'anàlisi exhaustiva per Sundgren (1991), qui ha elaborat una pro-

[13] Vegeu Sundgren (1994) (reproduït en aquest mateix número de **Qüestiió**) i altres papers del mateix autor, per una classificació sistemàtica i exhaustiva de la metainformació referent a les micro-dades. Aquí només pretenem il·lustrar la problemàtica amb alguns exemples.

posta ambiciosa i maximalista, basada en la consideració de les macrodades a la llum del procés d'agregació de microdades, del que són resultat. La pregunta que tracta de respondre Sundgren és la següent: Quines descripcions d'unes macrodades determinades cal que acompanyin aquestes per tal que els usuaris potencials d'una taula (aïllada o formant part d'una base de dades) estiguin en situació de jutjar la seva utilitat i fiabilitat en un context d'ús determinat?

La resposta és, aparentment, molt simple. Cal donar informació, tant sobre la població objecte d'interès i les seves característiques (variables) com sobre els mètodes i processos que poden condicionar la fiabilitat dels resultats finals; és a dir, els processos de recollida de les microdades i els de la seva agregació per obtenir les macrodades.

D'aquí es dedueix (i Sundgren ho recorda expressament) que la metainformació a donar sobre unes macrodades ha d'incloure també la informació pertinent sobre les microdades de les que provenen i sobre el procés d'agregació que les relliga a ambdues.

En la seva proposta, Sundgren especifica que cal proporcionar, per a unes macrodades determinades, les descripcions següents:

a) Els paràmetres dels quals, per a cada població i domini d'interès, les macrodades es suposa que són estimacions.

b) Les variables, els objectes unitaris de les quals han estat investigats (observats) al nivell micro de cara a l'estimació dels paràmetres del nivell macro.

c) Complementàriament, la funció "ideal" d'estimació, per deduir els paràmetres a partir dels valors observats de les variables.

d) La manera concreta com han estat portades a terme les observacions, i totes les incidències susceptibles d'haver afectat la qualitat dels resultats.

e) La comparabilitat en el temps i en l'espai, en general i, específicament, pel que té a veure amb les classificacions i nomenclatures (nacionals i internacionals).

Des d'una altra perspectiva, i en un paper posterior, Sundgren (1993) afegeix consideracions particulars per a dues menes de macrodades diferents:

- les taules estadístiques publicades.
- els conjunts de macrodades emmagatzemades en suports electrònics o digitals (disquets, CD-ROMs, bases de dades interrogables en línia).

En aquest sentit, si bé la informació a subministrar és en essència la mateixa, el diferent nivell d'agrupament i forma de presentació exigeix/permet la preparació de peces de metainformació estadística molt diferents, per atendre/aprofitar les molt

diverses necessitats/potencialitats dels sistemes en qüestió (multiplicades pels nivells diferents d'exigència dels diferents usuaris: des del públic a l'especialista temàtic o estadístic)[14].

1.5. Heterogeneïtat de les peces de metainformació estadística

A l'extrema heterogeneïtat de les peces de metainformació estadística segons la seva *natura*, evidenciada en la descripció tipològica anterior, cal que hi afegim ara les heterogeneïtats derivades del *format* en què es presenten les peces de metainformació estadística, del suport en que es vehiculen, del context dels *sistemes* en què s'integren i de l'*objectiu* a què es destinen.

El problema del *format* n'és un de temible donats, d'una banda, la diversitat de sistemes informàtics existents, amb exigències/potencialitats diferents respecte del format de descripció de dades requerit/possible, i de l'altra les necessitats de coordinació internacional en un moment, com l'actual, en que convergeixen les potencialitats tecnològiques (xarxes telemàtiques, p. ex.) amb les necessitats polítiques (el fet de la Unió Europea, i la mundialització de l'economia, en general). Una part important de les preocupacions del *grup de treball METIS*[15] és justament la d'integrar les exigències de models (*IRDS = Information Resource Dictionary System o IRM = Information Resources Management*, d'ISO, etc.) i formats (EDIFACT, ODA, SGML, etc.) internacionals a les necessitats específiques de la metainformació estadística.

Un aspecte no banal del format és el referent a les alternatives de presentació de la informació digitalitzada: en imatge (mapa de bits), en ASCII, en format de processador de textos (Word, WordPerfect), general de consulta (PDF, d'Acrobat, p. ex.), en document composat d'alguna mena (OpenDoc, OLE, p. ex.), o bé en forma d'enregistrament estructurat d'una base de dades (o DBF, en general).

Heterogeneïtat també respecte del *suport*, una vegada més amb la doble visió de les exigències/potencialitats ofertes per cadascun d'ells, i dintre de la tendència a la generalització de l'edició electrònica en totes les seves formes.

Heterogeneïtat, igualment, derivada del context del/dels *sistema/es* en el/els que s'integren les peces de metainformació estadística, de manera especial en la seva in-

[14] L'expressió dual "necessitat/potencialitat" és emprada per tal de ressaltar que es tracta de dues cares de la mateixa moneda, segons que es miri l'aspecte restrictiu d'un sistema (les seves limitacions) o l'aspecte positiu (les seves potencialitats). En la mesura que els sistemes informàtics evolucionen històricament en el sentit de disminuir les limitacions i augmentar les potencialitats, és possible plantejar-se objectius (d'integració i de coordinació) més forts cada vegada, utilitzant sistemes més moderns i avançats. Tenint en compte que queda sempre el problema de la transició dels vells sistemes als nous.

[15] Ja esmentat a la nota 1). *Vegeu també l'apartat 2.4*)

terrelació. Així, per posar un exemple del que volem dir, una peça de metainformació estadística pot consistir en un text presentat en una finestra en la que diversos descriptors o "botons" (*hot spots*) són lligams (*links*) associats a d'altres peces d'informació, al si d'un sistema hipertext, de manera que l'usuari, com és típic d'aquests sistemes, pot navegar lliurement entre aquelles peces d'informació interrelacionades.

Aquesta situació és molt diferent de la que es produeix en sistemes de bases de dades relacionals, en les que les peces d'informació han estat, *ab initio*, formalment estructurades segons la lògica conceptual de la informació, traduïda en un determinat esquema d'entitats-relacions.

Heterogeneïtat, finalment, derivada de l'*objectiu* a què es destinen les peces de metainformació estadística. Un parell d' exemples il·lustraran a què ens referim:

a) En primer lloc, suposem que es tracti d'una *nota al peu d'una taula* o associada amb ella. Doncs bé, serà molt diferent si va adreçada a un informàtic (en el qual cas serà breu i molt codificada), a un temàtic (text tècnic rigorós i complet des del punt de vista del significat de les xifres), o al públic no tècnic (text explicatiu més llarg, i fins i tot pedagògic, però menys rigorós i sense detalls tècnics). De fet, una nota associada a una taula pot fer referència a les seves condicions de producció, a les de la seva reutilització en un context diferent, o simplement, a les del seu "bon ús" (per tal d'evitar males interpretacions del seu significat).

b) En segon lloc, suposem que es tracti d'una *tipologia de tabulació*, és a dir, el conjunt d'ítems que classifiquen una variable en una taula. En aquest cas, hem de recordar que tipologies molt diverses poden haver estat originades per la mateixa Classificació de base, segons ordres d'agregació diferents. Per exemple, l'activitat econòmica dels establiments pot classificar-se segons una tipologia a 4 sectors, o 24 branques, que són agregacions diverses dels ítems de la *Classificació Nacional d'Activitats Econòmiques*. Doncs bé, un problema típic amb el que es troba l'usuari d'informació estadística tabulada és el d'identificar l'abast dels ítems a què corresponen els rètols de files i columnes de la taula (molt abreujats normalment per raó de l'espai disponible) cosa que requereix notes explicatives a part. Però l'usuari pot voler a més reutilitzar les etiquetes normals de la tipologia en el seu context de treball, i això requerirà que el sistema de metainformació estadística ho prevegi. En un institut d'estadística és essencial que un únic sistema de Classificacions sigui la font d'autoritat de totes les tipologies que s'utilitzin, que aquestes li estiguin integrades, i així mateix, que siguin automàticament accessibles des de qualsevol taula que les utilitzi en algun dels múltiples sistemes difosos per l'institut (publicació electrònica, base de dades en línia). D'altra banda, l'usuari haurà de poder, a més, triar la llengua en què li interessi la tipologia en qüestió.

1.6. Escenaris d'usos i usuaris

La consideració de l'objectiu en vistes del qual els usuaris necessitaran accedir a peces de metainformació estadística ens porta a plantejar el tema dels diferents tipus d'usuaris d'un sistema de metainformació estadística.

Des d'un punt de vista general, poden ser usuaris de metainformació estadística tant els que utilitzen informació estadística per ella mateixa com els que la produeixen, retrobant així el doble punt de vista amb el que podem enfocar la metainformació estadística, i al que al·ludíem a l'inici d'aquest article. Així, usuaris i productors d'informació estadística són les dues categories bàsiques d'usuaris de metainformació estadística. Tal com ja hem apuntat, "productors" són els diferents tècnics i especialistes (estadístics, temàtics, informàtics) que treballen al si d'un institut d'estadística en tasques relacionades amb la producció d'informació estadística, i "usuaris" seran les persones i institucions que la utilitzen (tant microdades com macrodades), normalment externs a l'institut. Uns i altres necessitaran accedir a certes peces de metainformació estadística relacionades amb la informació estadística amb la que treballen o que volen utilitzar. I ja hem fet notar que la metainformació estadística necessària vindrà determinada pel tipus específic d'objectiu o tasca en el context del qual l'usuari de metainformació estadística la necessita. És tot el problema dels *usos*, que s'afegeix a la categorització dels *usuaris*, i que fa que la identificació de les peces de metainformació estadística necessàries per a cadascun d'ells sigui a la vegada més complexa però també més pragmàtica, en la mesura en què, si es posa en relació amb el sistema de metainformació estadística l'eficàcia d'aquest, en la pràctica la seva utilitat serà més gran.

En aquest punt, voldríem fer tres observacions:

a) La *primera* té a veure amb la funció de difusió de la informació dels instituts d'estadística, cada vegada més important, a la que ja hem al·ludit a l'apartat 1.2), i que exerceixen diferents menes de tècnics (als estadístics, temàtics i informàtics s'afegeixen els especialistes en sistemes d'informació i documentació). Els *difusors* d'informació estadística (a no confondre amb els tècnics de la unitat orgànica de Difusió, quan aquesta existeix) han passat de les seves tasques tradicionals d'editar les publicacions estadístiques i prestar serveis de biblioteca i documentació al disseny i implementació de tota una panòplia de sistemes d'informació, des de bases de dades bibliogràfiques o factuals interrogables en línia a diferents formes d'edició electrònica). Així, doncs, en la mesura en què conjunts de peces de metainformació estadística han de ser elements constitutius importants d'aquests sistemes, en relació amb la informació estadística que difonguin, els *difusors* dels instituts d'estadística es constituiran en una nova categoria d'usuaris de metainformació estadística.

b) La *segona* està relacionada amb l'anterior. En efecte, en la mesura en què els sistemes de difusió d'informació estadística incorporen peces de metainformació estadística, els tècnics que participen en la seva implementació podran integrar-hi les peces de metainformació estadística que ja existeixin i estiguin disponibles; d'altra banda, però, serà necessari que en produeixin moltes altres, bé perquè les existents no s'adaptin al context dels usuaris dels sistemes, bé perquè es tracti de peces de metainformació estadística no necessàries per a la producció, bé perquè estiguin adreçades a categories específiques d'usuaris o a necessitats molt concretes.

c) La *tercera* consisteix simplement a esmentar unes situacions concretes en les que es poden trobar diferents tècnics d'un institut d'estadística i que constitueixen il·lustracions de la varietat d'usos de la metainformació estadística en diferents contextos. Els exemples han estat suggerits per les funcions de persones reals de l'*Institut d'Estadística de Catalunya*, si bé hem substituït els seus noms per lletres, i estereotipat en certa manera les seves funcions:

> El tècnic ***A***, en la seva feina de satisfer demandes del públic de taules estadístiques específiques per ser transmeses en disquet, necessita recollir la metainformació estadística necessària, en forma de definicions i notes (que poden trobar-se a la base de dades BEMCAT, però també en d'altres llocs i en formats diversos).
>
> El tècnic ***B***, en la seva feina de constituir arxius de microdades (prèviament preparats per no infringir el secret estadístic) per a la seva difusió, necessita integrar també al suport electrònic escollit (cinta o CD-ROM) la metainformació estadística necessària: definicions de variables, segur, però també diferents aspectes de la producció, de cara a proporcionar elements per a jutjar de la fiabilitat i comparabilitat de les dades en diferents contextos, i també afegirà descripcions del mètode utilitzat per obtenir l'arxiu susceptible de ser difós com a subconjunt de l'arxiu individualitzat de base. L'arxiu de Documentació Tècnica (definit a la *Norma 0*, però no implementat) li podria proporcionar la informació necessària, però l'única manera de que no hagi de recopiar-la (o, pitjor, reredactar-la) seria que estés prevista aquesta necessitat i es redactessin i s'emmagatzemessin les peces corresponents en el moment de constituir l'arxiu (digital) de Documentació Tècnica de l'operació estadística en qüestió.
>
> ***C***, especialista temàtic/a en la preparació dels Padrons Municipals d'Habitants 1996, necessita consultar informació metodològica general (internacional i pròpia) relativa als Censos de Població anteriors.
>
> ***D***, informàtic/a en la seva feina de produir, a partir dels arxius de base, noves taules per a formar part de publicacions o per a integrar-les a la base

de dades BEMCAT, necessita, d'una banda, codis i paràmetres relatius als arxius per a l'ús correcte dels programes informàtics i, de l'altra, els literals per a etiquetes de files i columnes i notes textuals (definicions, comentaris...). Es barregen, doncs, necessitats internes molt tècniques i les externes, (textos per als usuaris), en el context de la base de dades.

E, estadístic/a en el transcurs del disseny operatiu d'una nova operació estadística, necessita incorporar-hi una tipologia pròpia, obtinguda (s'entén automàticament) per agregacions de classes i/o ítems d'una classificació oficial.

F, estadístic/a en el transcurs del disseny operatiu d'una nova operació estadística, necessita informació diversa (metodològica, definicional, tipològica i estadística) per tal de comparar l'eficiència de mostres diferents del mateix univers utilitzades en diferents operacions del passat.

G, per a preparar els originals de les publicacions estadístiques en el sistema Macintosh d'edició electrònica (*desk top publishing*) utilitzat, necessita manipular els paquets d'informació estadística i textual associats que rep, i en els formats adequats per minimitzar la feina.

H, en la seva feina de construir un CD-ROM per difondre informació censal, té moltes necessitats de metainformació estadística com, per exemple:

- recollir tota mena de definicions, notes d'ús, qüestionaris, etc.
- integrar etiquetes de tipologies de variables als rètols de files i columnes

Això implica, en general, que les necessitats derivades dels sistemes, directes o indirectes, de difusió (BEMCAT, CD-ROMs, publicacions, videotex, etc.) han d'estar previstes ja a les fases de la producció i articulació de sistemes de producció, so pena d'haver de reescriue i adaptar les mateixes informacions en diferents contextos.

Finalment, ***I***, un administrador de l'institut, necessita recuperar informació de calendaris, personal, costos, etc. d'una operació anterior per a comparar-los amb una altra anàloga en curs.

Etcètera.

Des de la perspectiva diferent d'identificar els usos a què un sistema de metainformació estadística hauria d'atendre de cara al seu disseny, Sundgren (1992) defineix 5 escenaris d'ús:

- Escenari 1: Perspectiva orientada a l'usuari final.
- Escenari 2: Perspectiva orientada a la producció.

- Escenari 3: Perspectiva orientada al disseny (per fases).
- Escenari 4: Perspectiva gerencial.
- Escenari 5: Perspectiva tècnica.

Per a cadascun d'aquests escenaris, Sundgren, analitza les tasques i subtasques relacionades, les necessitats de cada fase o les dels subsistemes implicats, aportant una sèrie de consideracions molt suggestives.

1.7. Cap a sistemes integrats de metainformació estadística

Arribats a aquest punt, creiem necessari explicitar un parell d'aspectes importants, que poden haver passat desapercebuts i que, en tot cas, no hem justificat a bastament.

1.7.1. La generació de les peces de metainformació estadística

El primer d'aquests aspectes es refereix a la responsabilitat de fabricació de les peces de metainformació estadística al si d'un institut d'estadística.

En efecte, hem parlat en algun moment donant per suposat que les peces de metainformació estadística existien prèviament, o es fabricaven com a part dels processos normals de producció d'informació estadística. La realitat és, però, que en gran part caldrà fabricar-les expressament, atenent als usos a què vagin destinades i (molt important) amb el format i les característiques formals requerides pel sistema o subsistema en el què hagin de ser incorporades.

Ara bé, un problema especialment delicat és el de a quina unitat orgànica o funcional atribuir aquesta responsabilitat, al si d'un institut d'estadística. D'una banda, són els tècnics de producció els que tenen la informació i els coneixements necessaris per produir les peces de metainformació estadística tècniques i d'ús relacionades amb la creació i manipulació de la informació estadística en general, sobretot les microdades, però també, en part, les macrodades. El problema és que això desborda el seu objectiu funcional i el seu interès i, en la pràctica, la documentació que estan ben disposats a generar és solament la que necessiten per a l'eficàcia de la seva feina. D'altra banda, els tècnics relacionats amb els mitjans i sistemes de difusió (en sentit ampli) són els que són sensibles per la seva funció a les necessitats dels usuaris i, per tant, a ells pertoca fer els esforços que calgui per crear les diverses i nombroses peces de metainformació estadística necessàries. Per fer-ho, però, els caldrà disposar de la informació, bruta per dir-ho d'alguna manera, a partir de la qual,

mitjançant manipulacions i tractaments diversos puguin crear-les; i això de la manera més automàtica que sigui possible.

A més a més, hi ha conjunts de peces de metainformació estadística que, essent necessària la seva disponibilitat o organització en sistemes de recuperació d'informació de cara als usuaris externs, no provenen de les tasques de producció. Per exemple, la legislació i la documentació tècnica metodològica i internacional.

Altres, encara, tot i estar relacionades amb la feina de producció, són de caire no tècnic com, per exemple, les que informen sobre les fonts (impreses o digitals) a través de les quals es difonen els resultats de les operacions estadístiques.

Altres peces de metainformació estadística, finalment, tot i ser eines de producció, tenen un caràcter d'ús general i polivalent, que obliga a articular-les en sistemes autònoms i complexos. Aquest és el cas de les Classificacions, Nomenclatures i Codis territorials.

Tots aquests conjunts addicionals de metainformació estadística poden ser també responsabilitat funcional dels tècnics de difusió (que, com hem dit, van més enllà de la unitat de Difusió, quan existeix).

1.7.2. L'organització d'un sistema de metainformació estadística

El segon aspecte es refereix a l'organització dels conjunts de peces de metainformació estadística en sistemes d'informació.

Fins ara, hem parlat de metainformació estadística o de sistemes de metainformació estadística indistintament o, en tot cas, sense justificar la diferència.

Ara bé, en introduir el concepte de "sistema" i parlar de "sistema de metainformació estadística", entenem que els diferents conjunts i les peces de metainformació estadística mateixes s'articulen en una unitat d'ordre superior, en la que les interrelacions entre elles i amb els outputs i prestacions oferts als usuaris prenen un nou significat a la llum de la nova globalitat. Segons la proposta METIS, elaborada per Prazenka (1994):

> *A Statistical Metainformation System (SMS) is the information system that uses and stores statistical metadata and produces statistical metainformation for purpose of supporting decision making concerning an information system. The object of the Statistical Metainformation System is the statistical information system.*

De fet, l'organització de la metainformació estadística en un o diversos sistemes, i la natura d'aquests, com també el seu entorn informàtic, és un tema subjecte a

debat. Si l'especificitat d'alguns conjunts de metainformació estadística (bé sigui per la seva natura, bé per la mena d'usuaris o usos a qui vagin destinats) reclama un tractament autònom (com és el cas de les Classificacions, com exemple paradigmàtic), les interrelacions entre els diferents sistemes resultants reclamen una coordinació global forta (de cara a minimitzar duplicitats i a maximitzar-ne l'ús eficient per part dels usuaris).

Avui dia, amb el rerafons de la revolució tecnològica en curs derivada de les transformacions de les tecnologies de la informació i la comunicació, la *integració* (paraula-clau del moment) de sistemes a múltiples nivells d'informació és a la vegada un objectiu ambiciós, però a l'abast.

Dit això, ¿com compaginar el llast que la història ha deixat en els instituts d'estadística en forma de configuracions informàtiques determinades (si no obsoletes, sí corresponents a concepcions diverses) amb la necessitat de caminar cap a sistemes veritablement integrats entre les diferents menes de metainformació estadística, les necessitats dels usuaris, i entre la metainformació estadística i la informació estadística associada?

La resposta no existeix encara, almenys suficientment contrastada i assumida. Però el que sí podem fer és passar revista a les diferents formes com, en la pràctica, els instituts d'estadística d'altres països aborden el problema, i també quins són els seus plantejaments i els dels organismes internacionals sobre la metainformació estadística. Aquest és el significat de la segona part d'aquest article.

2. LA DIVERSITAT DE L'EXPERIÈNCIA INTERNACIONAL

2.1. Introducció al panorama internacional

El mètode que utilitzarem per passar revista a l'experiència internacional sobre la metainformació estadística no és el de fer-ho tractant d'esbrinar com s'ho ha plantejat cada institut d'estadística, sinó el de veure, per a cadascuna de les possibles formes que pot adoptar el sistema global de metainformació estadística o cadascun dels components o mitjans autònoms de metainformació estadística, quins són els problemes específics que es plantegen i quines han estat les solucions que, en la pràctica, han estat proposades, mitjançant projectes o experiències.

D'aquesta manera, el que pretenem és, prenent les experiències esmentades com a simples il·lustracions de solucions o plantejaments, donar una visió més detallada de

la problemàtica implicada en els sistemes de metainformació estadística, en la mesura que els instituts d'estadística se la plantegen.

No cal buscar, doncs, el plantejament específic de cada institut d'estadística; només són esmentats aquells que aporten algun projecte o experiència en un tema concret, que hem conegut i ens ha semblat il·lustratiu.

D'altra banda, tampoc ens ha preocupat massa si els exemples aportats podien estar desfassats (de fet, gran part de la informació prové del *Statistical Metainformation Systems Workshop* de 1993, i també de les reunions del grup de treball METIS), entenent que el que ens interessa aquí és més exemplificar tipus de solucions que donar notícia de solucions "al dia".

Així mateix, reconeixem un cert desequilibri material en el tractament dels diferents punts, inevitablement supeditat a la informació disponible.

En definitiva, passarem revista als punt següents:

- Sistemes generals de metainformació estadística, orientats a la producció.
- Paquets de metainformació estadística especialitzats.
- Sistemes generals de metainformació estadística, orientats als usuaris.
- Sistemes documentals i metainformació estadística.
- La metainformació estadística com informació associada a conjunts d'informació estadística, en suports digitals de difusió.
- El sistema de metainformació estadística com instrument per a la coordinació.
- La metainformació estadística integrada en un sistema expert.

2.2. Projectes i experiències en curs

2.2.1. Sistemes generals de metainformació estadística, orientats a la producció

La primera opció a considerar consisteix a aplicar a la metainformació estadística la mateixa tècnica de construir bases de dades relacionals sobre grans ordinadors (*mainframes*), utilitzada generalment als instituts d'estadística per mantenir centralment algunes o totes les tasques de la producció estadística. Es tracta, doncs, de dissenyar sistemes generals orientats a la producció, en els que la metainformació estadística considerada és la d'interès per als tècnics involucrats en la producció, l'objectiu dels quals, implícitament o expressa, és la d'arribar a constituir sistemes

complets i *actius* (anomenant així els sistemes que pretenen no solament referenciar sinó també integrar la metainformació estadística amb la informació estadística associada al si del mateix sistema).

En la pràctica, les dificultats tècniques derivades de la mateixa ambició de l'objectiu, i també les resistències humanes a les fortes exigències de coordinació implicades (a l'haver de documentar amb més amplitud de la que cadascú requeriria per a les seves pròpies necessitats), fan que les aplicacions de metainformació estadística acabin essent en realitat parcials, d'una banda, i aïllades del sistema d'informació estadística, de l'altra.

2.2.1.1. L'òptica "Dictionnaires de Données Statistiques (DDS)" de l'INSEE

L'enfocament "diccionari de dades" (*data dictionary*), en el context dels sistemes informàtics en general, fou impulsat per Shoshani, i va donar lloc a la incorporació de programari especialitzat als entorns de bases de dades, com és el cas del *CDD* (*Common Data Dictionary*) de Digital, sistema orientat a l'objecte de repositori de dades que centralitza les metadades corresponents a les dades dels diferents sistemes corrent en l'entorn de referència (VMS, p. ex.). És utilitzat a l'*Institut d'Estadística de Catalunya*.

L'INSEE (1990) va generalitzar el concepte per donar-l'hi un significat general dintre del Sistema Estadístic de França. La seva experiència, descrita per Lazarou (1993) és paradigmàtica, havent realitzat ja al 1983 un primer prototipus del sistema *DDS* (*Dictionnaires de Données Statistiques*). El seu objectiu final era molt ambiciós, puix que es tractava d'arribar a constituir un conjunt de sistemes articulats, que oferissin globalment l'accés al públic al conjunt del Sistema Estadístic de França, tot i que l'INSEE no tenia (ni té encara) competències sobre ell.

En la pràctica, la implementació s'ha reduït al mòdul "Production" (només de l'INSEE), composat dels submòduls "DDS-Enquêtes" i "DDS-Nomenclatures", que en la seva versió actual es configuren sota l'arquitectura client-servidor com a bases de dades relacionals sota ORACLE sobre un servidor UNIX, i accessibles des de clients PC/Windows en xarxa local, existint interfícies amb SAS. Les entitats (en un esquema d'entitats-relacions) considerades per a les Enquestes són: *Questionnaire, Question, Spécification, Programme, Variable, Nomenclature, Poste, Fichier, Tableau, Incident, Concept, Archive, Journal.*

L'institut d'estadística de Rumania està implementant aquesta versió del *DDS* per al control de la producció d'enquestes, com un component d'un sistema més general, en desenvolupament (Vg. Marina 1994).

D'altra banda, l'òptica "diccionaris de dades" fou el model conceptual de referència sobre el que es basava la nostra proposta d'estratègia *DIDAC (Diccionaris de Dades de Catalunya)* per a l'*Institut d'Estadística de Catalunya*, presentada al 1990 en un informe intern.

Altres instituts d'estadística han pres opcions semblants, menys ambicioses, com a simple ampliació de la seva experiència en bases de dades relacionals aplicades a la metainformació estadística, com, per exemple, l'*Instituto Nacional de Estadística* espanyol.

En efecte, l'INE (1994) (*Vegeu també Villar 1993*) que havia implementat al 1986 una petita versió d'un sistema de diccionaris de dades, ha desenvolupat al 1994 el *S.I.D. (Sistema de Información Documental)*, com l'estructuració informàtica mitjançant una base de dades relacional del conjunt d'informacions i documents tècnics relacionats amb cadascuna de les etapes de les operacions estadístiques a través del "Centro de Proceso de Datos", i d'acord amb la "Norma de Desarrollo de Aplicaciones", que paral·lelament s'ha implementat. Malgrat que es tracta, doncs, d'un sistema marcat pel seu entorn informàtic, en la mesura que en el seu disseny (via la *Norma* esmentada) s'han tingut en compte les informacions necessàries al conjunt de tècnics de l'INE, té un significat volgudament ampli, reduït tanmateix per l'orientació de producció inherent al plantejament.

2.2.1.2. L'enfocament "information management"

El plantejament "information management" que fa Ronald Graves (1994), en el que es basa l'actual estratègia de "Statistics Canada", respon a una consideració conceptualment simple i que té l'avantatge de permetre integrar els sistemes relacionals existents amb els nous tractaments documentals: la de considerar que en l'estadística, tenim dades i metadades. Mentre les dades (xifres, taules, arxius numèrics) són abundants i han estat tractades amb detall, les metadades (textos i documents de tota mena) estan disperses i han estat descurades. Tenim, doncs, dos paquets d'informació, que han de ser tractats amb mètodes diferents, constituint subsistemes d'un sistema superior que els integra i interrelaciona. Graves, d'una banda, estableix un esquema conceptual del sistema global, a través d'una definició operativa d'*informació* (que entén englobant *dades* i *documents*) i, de l'altra, preveu, al costat de les bases de dades relacionals que tracten les dades, la utilització de sistemes capaços de tractar informació heterogènia (referències, textos, documents composats, etc.), que no són únics (sistemes documentals tradicionals, sistemes de gestió d'informació, com BASIS/Plus, sistemes hipertext, etc).

L'objectiu és el d'anar integrant sota l'esquema global diversos mòduls (com *IBOSS = Internal Bank of Statistics System i EBOSS = External Bank of Statistics Canada*) i també productes (el catàleg de fonts, CD-ROMs específics, i d'altres).

Aquest plantejament apareix com a resultat de l'anàlisi de l'experiència de "Statistics Canada" en l'elaboració de diversos productes relacionats amb la metainformació estadística, com el CD-ROM *CANSIM* (semestral): conté bàsicament 500.000 sèries temporals (actualitzades semestralment) i està provist d'una interfície de consulta i explicació en hipertext, articulada amb el *SDDS (Statistical Data Directory System)*, veritable sistema de metainformació estadística estructurat sota BASIS. D'altres CD-ROMs publicats fins ara són el LMAS (mercat de treball) i els censals.

Afegim que hi ha relacions evidents entre aquest plantejament i alguns dels que esmentarem a l'apartat 2.2.4) i al 2.2.6).

2.2.2. Paquets de metainformació especialitzats

Aquesta opció suposa la renúncia (expressa o, més sovint, implícita) a plantejar-se un sistema general de metainformació estadística, per a concentrar-se en el plantejament de solucions per a conjunts particulars de metainformació estadística.

És una opció molt estesa entre els instituts d'estadística, si bé amb característiques especials segons l'experiència de cadascun, i també segons la sensibilitat relativa als problemes de la difusió al públic respecte dels de la producció.

En general, són les Classificacions i Nomenclatures els tipus de metainformació estadística tractats prioritàriament, fins i tot en absència de la seva consideració explícita com a metainformació, simplement per la seva importància intrínseca, tant per a la producció com per a la difusió d'informació estadística.

Així s'està fent, per exemple, a l'*Institut d'Estadística de Catalunya*, que té en curs d'elaboració una base de dades (*SICONO = Sistema d'informació de Codis i Nomenclatures*), interrogable en línia sobre l'ordinador central VAX.

EUROSTAT, per la seva banda, que edita les Classificacions en disquets, entre d'altres suports, prepara el projecte EDINOMEN, independent però integrable amb DSIS, l'objectiu del qual és el d'establir un sistema d'intercanvi electrònic de Classificacions i Nomenclatures entre els instituts d'estadística (Lebaube 1993).

Esmentem, així mateix, l'existència de programari comercial, com p. ex. BLAISE, desenvolupat per al disseny, recollida i control de la informació d'enquestes assistits per ordinador. En la seva versió BLAISE III, el tractament de les dades i les metadades corresponents està integrat (Schuerhoff 1993). El *Centre for Educational Sociology* n'ha fet una aplicació (Lamb 1993), en el context del projecte EISI, del programa DOSES (EUROSTAT) (enfocat a sistemes experts en estadística).

2.2.3. Sistemes generals de metainformació estadística, orientats als usuaris

Aquesta opció és la resposta lògica quan la preocupació prioritària (no única) és la de proporcionar als usuaris l'accés a la metainformació estadística i les prestacions que responen a les necessitats, generals i les particulars de grups d'usuaris específics; essent l'objectiu del sistema el de permetre consultes intel·ligents, una recuperació d'informació eficient i un suport efectiu a l'ús de la informació estadística en els diversos contextos de treball dels usuaris.

En la pràctica, un sistema amb aquestes funcionalitats pot adoptar diverses formes, i consistir, bé simplement en una interfície d'usuari per accedir a bases de dades centrals o distribuïdes, bé en una veritable arquitectura client-servidor, bé en un sistema autosuficient, bé en alguna de les variants intermitges. Veiem-ne alguns exemples.

2.2.3.1. Interfície d'usuari integradora

Existeixen múltiples variants d'interfícies, més o menys mediatitzades per les característiques de les bases de dades a les que han de donar accés. En la mesura, però, en què es dona la importància que veritablement té al terme "interfície d'usuari", com a sistema que "tradueix" els conceptes en què els usuaris tenen plantejades les seves necessitats i usos potencials a l'estructura (conceptual, organitzativa, formal i material) amb què les peces de metainformació estadística estan emmagatzemades a les diferents bases d'informació disponibles, es van decantant elements comuns de solució, encara en estat embrionari.

1) **Tesaures**

Si es tracta de "traduir" conceptes, una eina tradicional és un *tesaure* estructurat, que organitza els termes constitutius d'un lèxic, juntament amb les seves interrelacions, incloent tota mena de sinònims i termes utilitzats en diferents contextos. Kopp (1993) descriu el Tesaure utilitzat per l'institut estadístic de Berlín i explica el seu rol integrador com a peça bàsica del sistema cooperatiu (interciutats) de metainformació estadística desenvolupat a Alemanya, sobre la base dels conceptes teòrics elaborats per Appel (1993). La raó d'utilitzar el tesaure com a la interfície d'usuari bàsica "*rau en el reconeixement de que només la integració del llenguatge corrent en un sistema de metainformació estadística pot reduir la complexitat, que seria massa gran en qualsevol altra circumstància d'assolir un projecte tan ambiciós*" (Appel 1994).

2) **Sistemes hipertext**

Una altra manera (no contradictòria amb l'anterior) és la d'enfocar la interfície com un *browser* d'un sistema hipertext; en aquest cas, el conjunt de mapes conceptuals o "cartes de navegar" constitutives del browser (juntament amb les funcionalitats

de navegació pertinents) poden limitar-se a donar accés a les peces de metainformació estadística emmagatzemades a les bases d'informació tradicionals, o bé (al menys en part) contenir les pròpies peces de metainformació estadística. Aquesta darrera opció és particularment interessant, quan es tracta d'informació no estructurada o sota formes gràfiques o emmagatzemades en mapes de bits (imatge).

De fet, el programari EMMA (*Vegeu més avall*) presenta un mapa conceptual en hipertext com una de les formes de consulta i ja hem citat el CD-ROM CANSIM, provist d'una interfície en hipertext. En realitat, si bé no hi ha encara massa exemples operatius, és una opinió generalitzada que els sistemes hipertext jugaran un paper important en les interfícies dels sistemes d'informació i metainformació estadística.

A l'*Institut d'Estadística de Catalunya* es va elaborar l'any 1989 *munCAT*, "Un hiperdocument d'estadística demogràfica de les comarques i municipis de Catalunya". Desenvolupat sota HyperCard per a Macintosh, es tractava d'un prototipus que implementava experimentalment una sèrie de conceptes originals constituint una veritable proposta d'interfície especialitzada: d'una banda, relligant una sèrie de taules dels Padrons Municipals d'Habitants 1986 amb conjunts molt diversos de metainformació estadística (notes, legislació, gràfiques, mapes automàtics de coropletes, bibliografia, etc), "navegable" mitjançant mapes conceptuals a diversos nivells; de l'altra, oferint una sèrie de funcionalitats, estructurades en un conjunt sistemàtic i coherent (*munCAT*, juntament amb d'altres realitzacions, està descrit per Canals 1992). Alguns d'aquests conceptes han estat aplicats a la interfície de consulta inclosa al CD-ROM *Cens de Població de Catalunya 1991*, en curs d'elaboració (desenvolupat sota FoxPro per a PC/Windows) (*Vegeu l'apartat 2.2.5.2*).

2.2.3.2. El sistema PC-DOK de "Statistics Sweden", orientat al document

Tal com explica Malmborg (1993), el plantejament del sistema PC-DOK respon, a la vegada, als plantejaments conceptuals i teòrics generats al si del mateix institut suec d'estadística, on treballa Sundgren, i a l'experiència dels sistemes ja existents:

- DOK, un sistema documental sobre mainframe, contenint la documentació tècnica sobre els sistemes de producció, relacionats amb el tractament informàtic.

- ARK-DOC, utilitzant el mateix sistema, però contenint descripcions d'arxius bruts d'enquestes (típicament en cinta magnètica).

- DAISY, sistema de metainformació estadística per difondre informació sobre micro i macrodades sobre el mercat de treball. Es tracta d'un sistema sobre PC i l'usuari selecciona i s'emporta jocs de disquets amb programari i una (meta)- base de dades.

PC-DOK ha de substituir DOK i ARK-DOC tot suplementant i ampliant DAISY.

Des del punt de vista conceptual, PC-DOK reposa sobre eines ja experimentades, com SCBDOK, normativa de descripció de la documentació tècnica d'operacions estadístiques de producció[16]. Des del punt de vista de disseny, PC-DOK, es basa en l'orientació a l'objecte, accepta documents composats, i conté una interfície gràfica d'usuari sota Windows, amb connexions via SQL amb les bases de dades centrals; un model gràfic de presentació de la informació estadística (Malmborg 1988) és part important de cara a la seva facilitat d'ús, i adopta una forma semblant a la d'un tesaure gràfic, interpretable també com un mapa conceptual navegable (hipertext).

Si bé el plantejament de PC-DOK és general (i és per això que l'hem inclòs en aquest apartat), combina tècniques tractades a d'altres apartats. Cal així mateix fer notar que el programari PC-AXIS, inclòs a l'apartat 2.2.5, constitueix un avenç de PC-DOK.

És interessant d'afegir que (segons un text intern de Per Cronholm, tramès per Malmborg via e-mail), al 1995 s'ha iniciat l'anomenat *Database project*, que està constituït per tres components: *Metadata*, *Micro/macro* i *Output*, que complementen i/o substitueixen els sistemes descrits suara.

Metadata tindrà 2 parts: una base de dades contenint textos i una altra contenint metainformació emmagatzemada en SQL. Explícitament, el seu objectiu és doble: el de ser la porta d'entrada per recuperar informació a les bases de dades i publicacions, i el de ser instrument de coordinació de la producció d'informació estadística.

El component en format SQL contindrà també les classificacions, nomenclatures i codis territorials. També contindrà les descripcions dels fitxers individualitzats (*observation registers*) i de les macrodades (*Micro/macro database*).

La base de dades *Output* organitzarà l'accés a tota aquella informació (micro, macro i meta).

2.2.3.3. *El sistema EMMA (Enhanced Metainformation Management Architecture), de World Systems (Europe)*

EMMA és el primer producte comercial de programari concebut específicament per suportar totes les necessitats d'un sistema global de metainformació estadística. Ha estat elaborat per World Systems (Europe) i dissenyat per De Vaney (1994c),

[16]Normativa que Sundgren reprén i re-elabora en les seves plantilles de les "Guidelines"; d'altra banda, l'Institut d'Estadística de Catalunya l'ha utilitzat com a referència per a la seva pròpia normativa de l'Arxiu de Documentació Tècnica.

mantenint un estret contacte tècnic amb el grup METIS; d'altra banda, el mateix De Vaney participa en diversos projectes relacionats amb la metainformació estadística (com els ja esmentats SMILE i EISI). EMMA, que té una arquitectura modular, està basat en els principis client-servidor i API (Application Program Interfaces) per als lligams amb d'altres aplicacions, està provist d'una rica interfície gràfica d'usuari, on utilitza tècniques hipertext, i està desenvolupat per a treballar sota PC/Windows.

La filosofia d'EMMA és la de construir un sistema de metainformació complet separat de les dades, que se suposa que resideixen prèviament en els sistemes d'un institut, dedicant, això sí, un dels seus mòduls (*Data Repository*) a la comunicació i accés amb les bases de dades, amb les que acaba configurant un entorn integrat. La seva primera versió, que funciona encara com a monousuari, està essent avaluada oficialment per diversos instituts d'estadística europeus.

Els mòduls bàsics són els següents:

1 *Subject-Matter Knowledge-Base Module (SMKB)*
2 *Metainformation System Module (MIS)*
3 *Nomenclature Management System (NMS)*
4 *Data Repository Module (DR)*

Als que s'afegeixen les funcionalitats necessàries per a l'administració i manteniment del sistema (*Vegeu les Figures 2 i 3.*).

2.2.4. Sistemes documentals i metainformació estadística

En el context de la metainformació estadística, en la mesura en què es tracta d'organitzar i integrar en un mateix sistema conjunts heterogenis de peces de metainformació estadística, l'opció d'utilitzar l'enfocament dels sistemes documentals és pertinent, tant si el que es fa *és* interpretar que el SME[17] és un sistema documental de ple dret, com si el que es vol és aplicar les tècniques documentals a conjunts aïllats de peces de metainformació estadística.

Prendre aquesta opció suposa al mateix temps que el SME és un sistema *passiu*, és a dir, que es limita a referenciar les dades estadístiques, i es tradueix a aplicar les tècniques dels sistemes de recuperació d'informació (RI) a les peces de metainformació estadística, considerades com a documents. Aquelles consisteixen tradicionalment a emmagatzemar substituts dels documents (en forma de fitxes descriptives, amb referències i descriptors o resums, indicatius de la informació continguda) i construir mecanismes de recuperació i consulta d'informació, més o menys sofisticats.

[17] D'ara endavant, utilitzarem les sigles SME per a referir-nos a un Sistema de Metainformació Estadística.

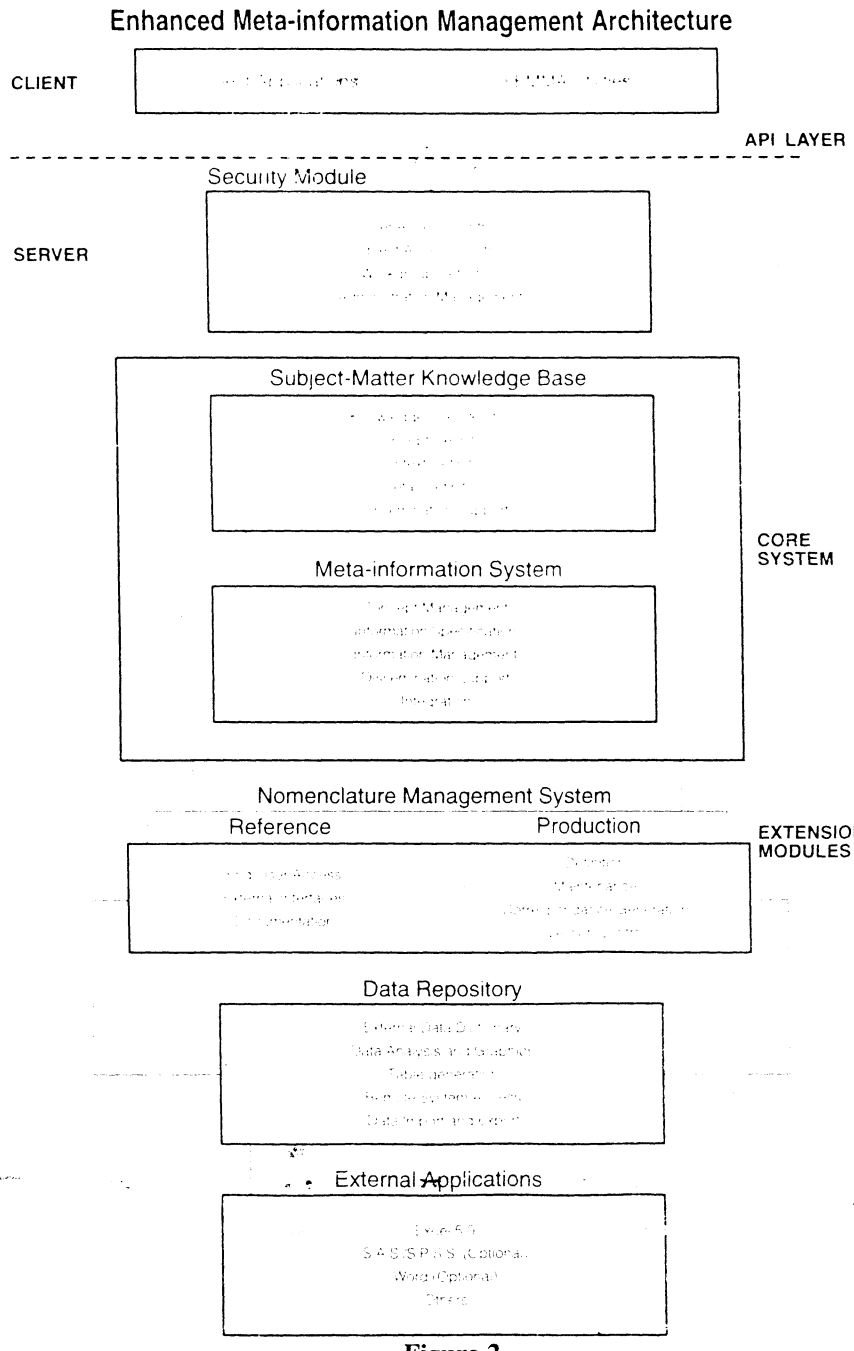

Figura 2

EMMA. Estructura modular.

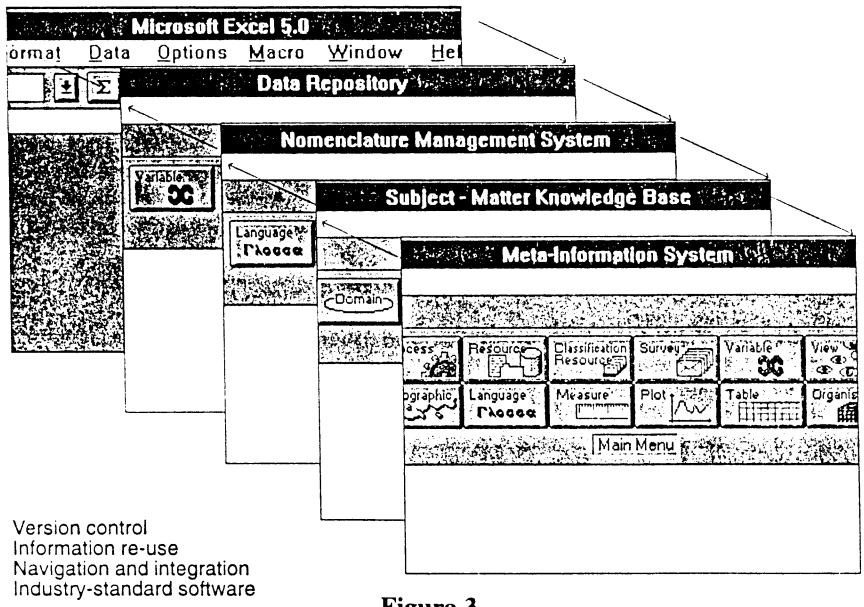

Figura 3

EMMA. Il·lustració de pantalles de mòduls.

Avui dia, però, els nous sistemes documentals es troben en plena evolució, en la mesura que els enregistraments poden contenir el text íntegre (en ASCII o PDF d'Acrobat, p. ex.) o, fins i tot, documents composats, en algun dels formats disponibles: CDA (DEC), Bento (OpenDoc,) OLE (Microsoft), etc. D'altra banda, les tècniques de recuperació s'allunyen de la mecànica de l'àlgebra booleana per adoptar algorismes més sofisticats (probabilístics, de proximitat semàntica, etc.)

Amb aquesta evolució es produeix una convergència dels sistemes documentals amb les bases de dades relacionals, però mantenint la més gran llibertat d'aquells en el tractament d'informació no estructurada.

Una alternativa radical a les tècniques de RI, cada vegada més sovint integrable a qualsevol sistema d'informació, és la de les tècniques *hipertext/hipermèdia*, que ofereixen la navegabilitat lliure entre les peces d'informació (il·limitadament heterogènies) com a forma de consulta específica, a base de seguir les associacions establertes (bé per l'autor, bé per l'usuari) mitjançant lligams (*links*) entre aquelles.

L'opció documental per als SME va normalment associada a un enfocament de la metainformació estadística orientada a l'usuari, en el que es prioritzen sobretot les necessitats dels usuaris externs. D'altra banda, la no-exigència d'informació estruc-

turada per part dels sistemes documentals els dota d'una gran capacitat integradora de sistemes de tota mena, especialment bases de dades heterogènies.

Ja hem pogut veure l'aplicació de tècniques documentals en alguns dels sistemes esmentats fins ara (l'esquema de Graves per al Canadà, PC-DOK, munCAT), i també la veurem en d'altres dels que ens ocuparem en els apartats següents.

Per acabar, només volem esmentar una aplicació específicament documental: la dels sistemes que donen informació sobre les fonts estadístiques rellevants, que hem identificat a l'apartat 1.1) com una de les menes de metainformació estadística més importants per als usuaris finals. No debades Allin (1993) identifica aquesta necessitat d'informació com la primera per a aquests, d'entre les 7 capes en les que agrupa les necessitats de metainformació estadística dels usuaris finals.

Malgrat això, són encara pocs els països que disposen d'una "guia de fonts estadístiques", la necessitat de la qual és cada vegada més reconeguda. El mateix Allin (1993), per exemple, en reclama una per al Regne Unit. I, en l'esquema d'"Statistics Canada", està prevista la conversió del "Statistics Canada Catalogue" en una base documental informatitzada.

Fou l'INSEE el primer que va publicar el "Répertoire de Sources Statistiques" (als anys 60s) i el Servei d'Estudis de Barcelona del Banc Urquijo qui publicava al 1969 la primera "*Guía de Fuentes Estadísticas de España*", continuada i ampliada pel *Consorci d'Informació i Documentació de Catalunya* amb l'*Inventario de Estadísticas de España*, complementat després amb les dues bases de dades públiques (sota BASIS) interrogables en línia: ESPAN i ESCAT (aquesta sobre les fonts estadístiques de Catalunya). Actualment, estan sotmeses a revisió sobre la base de les competències que la llei atribueix a l'*Institut d'Estadística de Catalunya* per a coordinar els organismes productors d'estadística.

2.2.5. La metainformació estadística com informació associada a conjunts d'informació estadística, en suports digitals de difusió

Un problema especial de tractament de la metainformació estadística és el que es planteja quan un institut d'estadística, dintre de la seva estratègia de difusió general, es proposa incorporar la metainformació estadística associada als conjunts d'informació estadística que difon al públic. En realitat, no es tracta d'un sol problema sinó de tants com mètodes i mitjans de difusió utilitzi, i la veritat és que, avui dia, a més del mitjà clàssic de les publicacions impreses (encara, naturalment, indipensables, si bé no sabem per quant de temps), l'oferta tecnològica i metodològica és variada i en curs de transformació, basada, naturalment, en la digitalització prèvia de la informació. És el que s'ha anomenat l'edició electrònica.

Una classificació operativa d'aquests mitjans pot ser la següent:

1) Disquets.
2) CD-ROM i altres suports de memòria massiva.
3) Bases de dades interactives públiques.
4) Servidors *WWW (World Wide Web)* per Internet.

2.2.5.1. Disquets

Disquets contenint taules preformatades és una manera simple i barata de distribució d'informació estadística, oferint una gran varietat de possibilitats: des de taules en ASCII (amb delimitador de columnes pel tabulador) fins a sèries de taules o arxius de microdades provistos de programari especial per a l'extracció, amb possibilitats de certa manipulació de taules.

Aquesta darrera forma és la més útil a l'usuari, però exigeix haver definit algun format que permeti la seva reutilització per a diferents informacions, i que faciliti els intercanvis d'informació entre un institut i els seus usuaris. Exemples interessants de formats incorporats a programari: *STATVIEW* (Netherlands Statistical Institute), *EXTRACT* (US Bureau of the Census) (Zeisset 1993) i *PC-AXIS* (Statistics Sweden) (Nordbäck 1992).

L'INE, per la seva banda, ha desenvolupat el programari *INEDAT*, per a PC/MS-DOS, que permet visualitzar i extreure taules i documents a partir de les seves publicacions estadístiques en disquet. Una nova versió gràfica permet visualitzar informacions en finestres simultànies, i accedir des d'una taula a la informació metodològica associada. També distribueix en disquets algunes Classificacions, provistes d'un sistema de consulta per a PC/MS-DOS (*SIN=Sistema Informatizado de Nomenclaturas*).

El problema general és com incorporar la metainformació estadística, donada l'escassa capacitat dels disquets quan hi volem incorporar informació en formats rics (la qual cosa obliga sovint a comprimir- la). Una possibilitat consisteix a difondre disquets on les taules i la documentació annexa només poden ser consultades, sense possibilitats de manipulació per l'usuari. S'utilitzen aleshores llenguatges orientats als objectes com C++ o programari de desenvolupament hipertext (HyperCard, ToolBook, Folio, Smartext, etc.). Exemples recents en són:

Bizkaiko Zentsuak (Censos de Bizkaia). Diputació Foral de Bizkaia.

Informació de Base. Pla Territorial Metropolità de Barcelona (Generalitat de Catalunya)

Catalunya en Xifres. Institut d'Estadística de Catalunya

2.2.5.2. *CD-ROM com a suport de memòria massiva*

La gran virtut del CD-ROM és la seva gran capacitat de memòria, que permet eixamplar enormement les possibilitats d'emmagatzemament, gestió, manipulació, presentació i extracció, tant de la informació estadística com de la metainformació estadística associada, per ser utilitzat mitjançant aparells lectors de CD-ROM en entorns PC o Macintosh, en general.

No tot, però, és possible alhora i, en l'estat actual de les coses, un compromís és encara necessari entre la sofisticació i multiplicitat de prestacions i la simplicitat d'ús per a l'usuari, en un context de demanda potencial real encara reduïda (tot i que creix ràpidament).

No podem aquí detallar la problemàtica general de la informació estadística distribuïda en CD-ROM. Deixem constància només de la diversitat de plantejaments dels instituts d'estadística, que utilitzen criteris, metodologies i programari diferents, la qual cosa representa un problema afegit per als usuaris.

Respecte de la metainformació estadística, comencem per dir que en el CD-ROM ja no tenim els problemes de capacitat a que al·ludíem al parlar dels disquets. Així podrem, en principi, pensar en incorporar informació textual en formats tipogràfics rics, i també gràfiques, o mapes, per exemple, i utilitzar-los no simplement com a peces d'informació passives, per a ser desplegades a la pantalla, sinó també per ser manipulades per l'usuari o bé formant part d'interfícies gràfiques d'usuari completes.

El problema n'és un de metodològic i tècnic en el context dels sistemes d'informació. D'una banda ¿com associar cadascuna de les peces de metainformació estadística a la informació estadística corresponent (i recordem l'heterogeneïtat d'aquella)? De l'altra, com construir al mateix temps i, complementàriament, un sistema de consulta de la metainformació estadística prou sofisticat per ser útil a usuaris molt diversos (en les seves necessitats i en els seus coneixements)?

La solució passa per fer compatibles i coherents en un mateix sistema dos sistemes de filosofia diferent: l'enfocament *base de dades*, adequat per a tot el que fa a la informació estadística, i l'enfocament hipertext, adequat pel que fa a les peces de metainformació estadística i a la interfície general d'usuari. Així ho van descobrint amb la pròpia experiència els instituts que han elaborat CD-ROMs, tot i que l'estat de la tecnologia no permet desenvolupar fàcilment aquestes solucions híbrides, més fàcils i potents ara com ara en l'entorn Macintosh que sota PC/Windows. De tota manera, l'evolució general cap a sistemes orientats a l'objecte promou la convergència tècnica entre aquells dos enfocaments; a més a més, en la pràctica actual, cada vegada hi ha més casos d'incorporació de tècniques hipertext als sistemes clàssics de bases de dades. El *Help Compiler* i el *Multimedia Viewer* de Microsoft en són exem-

ples. Esmentem també *LinkWorks*, de Digital, que permet desenvolupar interfícies en Macintosh i PC/Windows associades a bases de dades corrent sobre VAX (en entorn VMS).

De fet, retrobem, encapsulats en un suport material, tots els problemes de la metainformació estadística, en general. L'avantatge que tenim ara és el de que, al tractar-se d'un suport autònom d'informació, el CD-ROM permet als instituts d'estadística experimentar en l'organització de sistemes de metainformació estadística a petita escala, sense interferir amb la producció, tenint a més la possibilitat de l'efecte demostració sobre l'eficàcia de certes tècniques.

Tanmateix, no cal oblidar que la informació i metainformació estadística a encapsular en un CD-ROM segueix essent tributària dels sistemes d'informació de l'institut i, per tant, caldrà preveure que subministrin les peces d'informació i de metainformació estadística *amb les característiques i en la forma i format específics al sistema utilitzat al CD-ROM*, sota pena d'haver de repetir molta feina inútilment. D'aquí se'n dedueix que, si bé el CD-ROM permet construir sistemes autònoms, el seu plantejament no pot portar-se a terme totalment a banda dels sistemes de producció; cal un plantejament global, en el que el CD-ROM sigui una peça del conjunt. És més, partint de la base que un nombre a priori indeterminat, però important, de peces de metainformació estadística caldrà que siguin definides i elaborades expressament per al CD-ROM, donat que no corresponen a necessitats del sistema de producció, serà convenient que l'esmentada definició sigui feta al si del mateix plantejament global, preveient que, en el futur, estigui compresa i sigui coherent amb l'estratègia de l'institut en qüestió.

Com a exemples de CD-ROM amb plantejament específic de metainformació estadística, podem reprendre, en primer lloc, els ja esmentats *LMAS* i *CANSIM*, del Canadà, descrits per Podehl (1993). El CD-ROM *LMAS (Labour Market Activity Survey)* conté informació provinent de l'enquesta sobre el mercat de treball a diferents nivells. La documentació sobre l'enquesta a donar als usuaris ocupa varis volums i és molt complexa, de manera que s'ha organitzat en un sistema hipertext, desenvolupat sota *Folio Views* per a PC/Windows, que complementa i s'articula amb un altre sistema de gestió de dades, desenvolupat amb un programari de la *Ohio State University*. El conjunt ocupa 2 gigabytes d'informació distribuïts en 3 discos compactes.

El CD-ROM *CANSIM*, que es publica semestralment des de fa uns anys com a actualitzacions de la base de dades de sèries temporals del Canadà no contenia fins al 1993 altra metainformació estadística que unes definicions bàsiques, moment en el que fou reelaborat per tal de, a partir de l'experiència de LMAS, incorporar-hi un sistema hipertext. Aquest consistia (segons el projecte de Podehl 1993) en articular una interfície entre el banc de sèries, el directori de dades (*SDDS=Statistical Data*

Directory System) i el catàleg "explicat" de publicacions, de manera que l'usuari hauria de poder anar de l'un a l'altre seguint el fil d'un tema determinat.

Un altre exemple d'articulació de la metainformació estadística en un CD-ROM és el *Cens de Població de Catalunya 1991* en curs d'elaboració a l'*Institut d'Estadística de Catalunya*, per ser consultat sota Windows. La informació estadística consisteix en un conjunt de taules bàsiques censals per a tots els municipis de Catalunya i tots els nivells territorials superiors. La interfície gràfica de consulta, desenvolupada sota *FoxPro/Windows*, permet seleccionar, manipular i exportar taules a diferents nivells, visualitzant-les en finestres simultànies, i obtenir automàticament piràmides de població, gràfiques i mapes de coropletes, mitjançant barres de menús i paletes de navegació (Vegeu Figura 4).

En el que fa a la metainformació estadística, s'han previst dos entorns diferents, utilitzant tècniques hipertext. Al primer, consistent en finestres especials d'*Info*, s'accedeix quan l'usuari prem la icona "i", i es presenten les definicions i tipologies de tabulació corresponents a la taula que està activa en aquell moment. Aquestes finestres d'*Info* contenen a més "botons" que permeten accedir a l'altre entorn, el mòdul META, anant a parar directament a peces de metainformació estadística relacionades amb aquella taula activa. Aquest mòdul és així un entorn connectat amb el de les taules, però autònom, i ha estat desenvolupat utilitzant el *Help Compiler*, integrable amb *FoxPro/Windows*, i que disposa de la seva pròpia interfície de consulta i navegació. La metainformació estadística continguda és la documentació metodològica bàsica del Cens de Població 91. Podria, eventualment, utilitzar-se el *Multimedia Viewer*, enlloc del Help Compiler, ja que es tracta del mateix entorn, però amb moltes més prestacions, i permetria construir un sistema de metainformació estadística complet a diferents nivells sobre les múltiples i heterogènies peces d'informació del *Sistema Estadístic de Catalunya*, com ja es va experimentar en el prototipus *munCAT*, sobre Macintosh.

Citem també el CD-ROM "Anuari estadístic de la ciutat de Barcelona. 1993", que és un reflex fidel de la versió en paper; es tracta, doncs, d'un sistema provist d'una interfície de presentació de textos i taules, sense possibilitat de tractament, desenvolupat sota Microsoft Viewer, i editat per l'Ajuntament de Barcelona.

Un altre ús del CD-ROM per part dels instituts d'estadística és el d'emprar-lo com a vehicle de difusió d'arxius de microdades. En aquest cas, es prioritza simplement la capacitat de memòria massiva d'aquell suport sobre la sofisticació de la interfície, i no s'hi inclou normalment cap mòdul de metainformació estadística, entenent-se que els usuaris d'aquesta mena d'informació són experts que coneixen a bastament amb quina informació se les han d'haver. Aquest ús del CD-ROM, suport de memòria òptica, l'assimila a d'altres suports de memòria magnètica massiva.

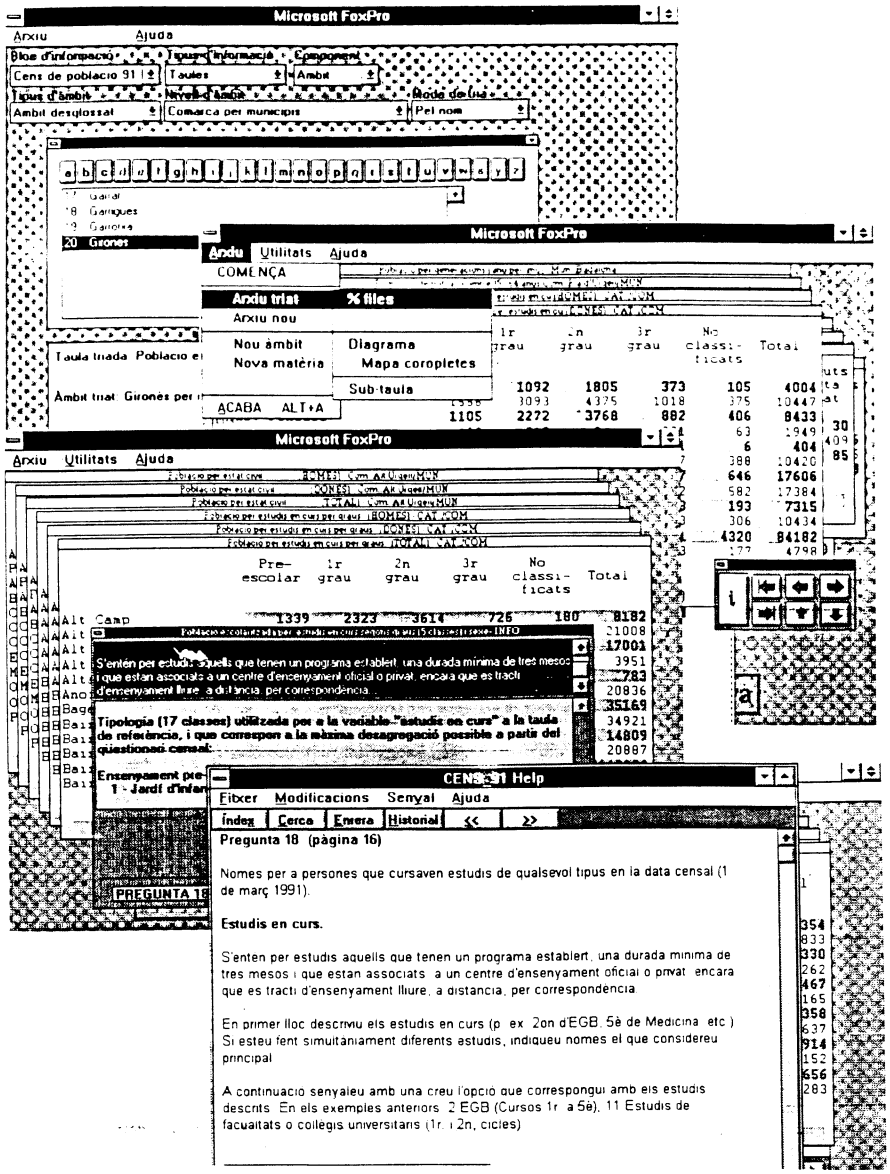

Figura 4

CD-ROM "Cens de Població de Catalunya 1991". Il·lustració de pantalles. Menús, finestra Info, Finestra Meta...

2.2.5.3. Bases de dades interactives públiques

En les bases de dades interactives que els instituts d'estadística han anat constituïnt i posant al servei del públic (directament, o a través de servidors comercials), tant si eren bases de dades bibliogràfiques (tipus "guies de fonts estadístiques") com bases de dades numèriques (tipus "sèries temporals" o "arxius estadístics de base territorial"), no es preveia en general cap mòdul ni tractament específic de metainformació estadística.

La raó és doble: d'una banda, el propi plantejament metodològic de la metainformació estadística és recent; de l'altra, hi ha dificultats tècniques per a engarçar i articul·lar la metainformació estadística amb el mode de tractament de la informació numèrica utilitzat per oferir la interrogació pública, que normalment correspon a desenvolupaments molt específics de l'enfocament bases de dades relacionals sobre grans ordinadors (mainframes), els quals només recentment han començat a incorporar lentament l'orientació a l'objecte. Justament, aquestes dificultats han fet que l'opció CD-ROM s'utilitzi com a terreny d'experimentació de noves solucions, tal com hem apuntat.

A més, l'òptica tradicional d'interrogació de bases de dades numèriques obligava els usuaris a aprendre i manegar llenguatges complexos de definició, selecció i extracció de taules, en un entorn relativament opac per als no experts, que ja ni intentaven aclarir-s'hi.

Deixem constància de que s'ha realitzat per part dels desenvolupadors un cert esforç per fer més còmode l'accés a les dades (tant si són arxius de sèries temporals com taules), a través de noves interfícies d'usuari on, malgrat tot, la documentació metodològica, quan hi és, està disponible en mòduls separats, si bé en certs casos s'hi pot accedir directament des d'una taula.

a) *BEMCAT i SAETA*

La base de dades *BEMCAT (Base d'Estadístiques Municipals i Comarcals de Catalunya)*, desenvolupada per l'*Institut d'Estadística de Catalunya* n'és un exemple. Conté taules amb estadístiques de tota mena relatives als municipis, comarques i altres nivells territorials de Catalunya, i dóna accés opcionalment a unes "descripcions temàtiques" associades a una taula o als nivells temàtics superiors. *BEMCAT*, base de dades relacional RDB residint en un ordinador VAX, és accessible (de moment, internament) bé per una interfície de menús (per emulació de terminal VT200), bé per una interfície gràfica en OSF/MOTIF.

Una altra forma de procedir consisteix a proveir els usuaris amb un programari de selecció i extracció de taules. Un exemple n'és el programari *SAETA (Sistema de Archivo y Extracción de Tablas)*, desenvolupat per l'INE per a visualitzar i

extreure taules i documents metodològics associats amb elles a partir de les seves bases de dades. També el PC-AXIS, de l'institut suec, ja citat, fa aquesta mateixa funció.

b) *MDB (U.S. Bureau of the Census)*

El plantejament de futur que fa l'*US Bureau of the Census (BOC)* respecte de la metainformació estadística confirma les dificultats per oferir solucions combinant en un sol sistema la informació estadística i els paquets de metainformació estadística necessaris. Tal com explica Gillman (1994), el BOC s'ha concentrat en treballar sobre la metainformació estadística en paral·lel a les bases de dades numèriques dissenyant, per encàrrec del Departament de Comerç, un prototipus de base de dades pública anomenada *MetaDataBase (MDB)*, l'objectiu de la qual seria el d'assistir als usuaris d'informació estadística en la localització dels conjunts específics de dades pertinents a les seves necessitats concretes, ocasionals o sistemàtiques. La informació estadística de referència prevista per a la base de dades final no es limita a la del BOC; al contrari, es pretén incorporar totes les fonts de dades de l'administració federal, si bé el prototipus inicial es limita a un subconjunt de la informació del BOC. D'altra banda, quedaria per a un projecte posterior l'articulació directa amb les dades estadístiques mateixes.

Als efectes de *MDB*, la metainformació estadística d'interès es divideix en tres categories: *de sistema, d'aplicació* i *administrativa*, per tal de satisfer les diverses necessitats dels usuaris. La *de sistema* es refereix, bàsicament, a informació sobre la lògica i el tractament informàtic utilitzat en la creació de les dades; la *d'aplicació* comprèn els aspectes estadístics necessaris per interpretar correctament les dades, i l'*administrativa* correspon a aspectes institucionals i administratius necessaris per accedir a les dades i a les persones de contacte pertinents. Com es veu, l'objectiu és a la vegada comprensiu i pràctic en el que fa a la metainformació estadística, si bé, recordem-ho, parcial, en la mesura que *només* intenta crear un SME deslligat de les dades mateixes.

El model de referència del sistema MDB en ell mateix és l'estàndar *IRDS (Vegeu l'apartat 2.3.3)* i es preveu connexions amb els sistemes de recuperació ja implementats i públicament disponibles per a conjunts particulars de metainformació estadística, com són, p. ex.: *ARRK (Automated Reference Rack)*[18], *DES=Data Extraction System* (Survey-on-Call) i *CENDATA* (sobre dades censals, accessible per Dialog i Compuserve), etc. D'altra banda, també s'utilitzarà el format "BOX file" ja existent com a llenguatge de descripció de taules, que permet incloure camps específics de metadades.

[18]ARRK és un sistema hipertext, desenvolupat amb el programari de desenvolupament *Smartext*, de Lotus.

Un dels requeriments de *MDB*, (amb exigències de llarg abast en el que fa a la coordinació entre organismes productors) que té el seu paral·lel en una de les preocupacions de METIS, és el d'elaborar un lèxic comú terminològic i un model de dades coherent amb la nova filosofia i consistent amb les heterogènies pràctiques dels diferents productors.

Des del punt de vista informàtic, per a la primera realització del prototipus, encara a petita escala (enfocat a la verificació de la consistència lògica i pràctica del model), s'utilitza *Infospan Open Repository*, l'únic (fins a la data) sistema de desenvolupament existent complint les especificacions *IRDS*.

c) *Videotex*

Apart de les bases de dades interactives clàssiques, com a sistemes de recuperació d'informació, hem de citar el sistema Videotex que, per les seves característiques i limitacions, no permet utilitzar-lo fins ara més que com a sistema de presentació d'informació i metainformació estadística molt sintètiques per al públic en general; o bé com a sistema de publicació electrònica de dades molt freqüents i poc detallades (exemple típic, l'IPC = Índexs de Preus al Consum). Entre d'altres, l'INE manté a Espanya una base per Ibertex i l'*Institut d'Estadística de Catalunya* posa ara al públic una primera base d'informació estadística, anomenada XIFRES, també per la xarxa Ibertex.

2.2.5.4. Servidors WWW (World Wide Web) per a Internet

Un dels desenvolupaments dels darrers anys conduents a fer més fàcil l'accès i la interrogació de les bases de dades interactives ha consistit en desenvolupar interfícies d'usuari per a entorns PC/Windows i Macintosh, i que actuaven bé com a interfícies de comunicació simples, bé com a clients en un entorn client-servidor.

Sense invalidar aquest desenvolupament, els instituts d'estadística més avançats no han deixat d'explorar les noves possibilitats d'accés pràcticament universal ofertes per la xarxa Internet, de primer constituint servidors d'arxius "ftp anonymous" i "gophers" i, darrerament, creant servidors de pàgines *World-Wide-Web (WWW)*, en el que està esdevenint una moda abassegadora o, potser, un nou servei indispensable per a distribuir no solament informació estadística (tant micro com macrodades) sinó també i, pel que a nosaltres interessa, sobretot, metainformació estadística. La raó és senzilla i se sustenta en una doble constatació: d'una banda, permet continuar posant a disposició del públic les bases de dades estadístiques ja existents, però amb la possibilitat de construir interfícies gràfiques d'usuari relativament independents dels sistemes del servidor i, al mateix temps, d'una gran simplicitat, amb l'avantatge addicional que, definint un sol entorn, es posa a disposició pràcticament de tot el món;

d'altra banda, la filosofia de base de l'entorn *WWW* és hipertext, i les pàgines web s'escriuen amb un llenguatge d'orientació "document composat" (*HTML=HyperText Management Language*, un subconjunt de l'estàndar *SGML*), que poden anar provistes de "botons" (*links*) que les associen no solament amb d'altres parts del mateix servidor sinó també amb les de qualsevol altre servidor WEB d'Internet. Les possibilitats que aquesta tecnologia ofereix per a la metainformació estadística són, així, impresionants, tot i recordant els problemes de tipus conceptual i d'estandardització als que hem fet referència més amunt.

Com explica Cathryn Dippo (1994), l'*US Bureau of Labour Statistics* aprofita l'entorn *WWW* per a distribuir la informació estadística (bàsicament, sèries temporals territorialitzades i enquestes especials) ja disponible a les seves bases de dades *BLS*, a través d'una interfície gràfica Web, que simplifica la selecció de taules i el format d'exportació desitjat. A més, el *BLS* s'ha pres seriosament l'oportunitat d'aprofitar les prestacions hipertext i la presentació de documents en format ric (PDF d'*Acrobat*) per oferir diversos conjunts de metainformació estadística: d'una banda, la documentació metodològica associada als arxius estadístics, cartogràfics i gràfics i, de l'altra, documentació metodològica general, programari especialitzat i informació institucional (l'actualitat estadística, organigrames i persones a contactar, calendaris d'operacions estadístiques, etc.).

2.2.6. El sistema de metainformació estadística, com a instrument per a la coordinació

Ja hem esmentat en algun moment que la metainformació estadística pot jugar un paper integrador de primer ordre, en la mesura que se situa en el punt més proper a l'usuari i, per tant, més global (*Vegeu, a més, l'apartat 2.4*). D'aquí a fer-li jugar el paper d'instrument per a la coordinació de diverses institucions no hi ha més que un pas.

Així ho ha vist, per exemple, l'institut d'estadística de Berlín que, tal com defensa Appel (1994), entén que un SME és una precondició per assolir una estandardització bàsica en els processos de producció i, així, estén el seu sistema *DUVA* a les 50 grans ciutats alemanyes, l'objectiu del qual és el de coordinar els registres municipals administratius de població. Ja hem esmentat a l'apartat 2.2.3.1 el paper fonamental que el Tesaure jugava al si del seu SME com un dels sis mòduls (Tesaure, Textos, microdades, macrodades, Directori i Confirmacions de qualitat). En l'articulació progressiva del SME de DUVA es van establint normes consensuades de procediments i definicions, entorn d'un model de dades, que incorpora expressament els estàndars internacionals, com els requisits GESMES (EDIFACT) per a l'intercanvi electrònic d'estadístiques.

Un altre exemple n'és *GENIE* (Walker 1993), un macroprojecte liderat per la Universitat de Loughborough que consistirà en el sistema de base del *Global Environmental Change Data Network*, que utilitza al seu torn la xarxa JANET.

L'esmentem aquí, encara que l'objectiu del sistema es redueix a recollir i proporcionar informació al servei dels centres de recerca en una temàtica específica (el canvi climàtic del món) i els productors d'estadístiques oficials només hi participen com a proveïdors de dades, per la raó que el nucli integrador de les bases de dades a constituir serà un sistema de metainformació estadística, la definició i alimentació dels elements del qual serà l'arma principal en la coordinació de les interrelacions amb els diversos centres.

Recordem, finalment, com el projecte DSIS d'EUROSTAT atorgava un paper integrador al *common reference environment*, del que la metainformació estadística és part fonamental.

2.2.7. La metainformació estadística, integrada en un sistema expert

Tal com explica Hand (1992), una de les àrees més prometedores de l'aplicació de les tècniques de la Intel·ligència Artificial (IA) a l'estadística és la dels sistemes experts, en la mesura que es proposen ajudar els usuaris no-experts en l'anàlisi de la informació estadística, suggerint, p. ex., hipòtesis, mètodes alternatius d'exploració i d'interpretació, etc., per a conjunts de dades concretes.

D'Angiolini (1993) va més lluny defensant que el propi sistema d'informació estadística pot enfocar-se amb l'òptica "coneixement", definint-lo aleshores com:

"el corpus de coneixements d'un agent (personal o institucional) l'objectiu del qual és augmentar el seu propi coneixement sobre l'entorn, mitjançant activitats estadístiques com:

- recollir informació (estadística) a base d'observar parts determinades del món real,
- aplicar tècniques d'anàlisi estadística per a augmentar el seu coneixement".

En aquest context, és evident que la metainformació estadística ha de jogar un paper fonamental en la configuració i contingut de la base de coneixement del sistema. El programa *DOSES* d'EUROSTAT així ho ha entès al seleccionar dos dels projectes especialment dedicats a la metainformació estadística: *MMD* i *EISI*.

MMD (Modelling Metadata), descrit per Darius (1993) consisteix en un model formalitzat de l'ús de la metainformació estadística en relació amb les diferents tasques dels usuaris d'un sistema d'informació estadística. Una anàlisi de les tasques des

del punt de vista dels coneixements implicats ha portat a la definició d'un conjunt d'operadors, articulables en cadenes, uns dels quals signifiquen manipulacions de les estructures de dades/metadades que donen com a resultat altres estructures de dades/metadades (p. ex, selecció d'una columna, fusió de conjunts, etc.); mentre d'altres generen resultats finals (generació de taules, generació de gràfics). Aquests operadors tenen com a característica especial que no es limiten a tractaments mecànics, sinó que tracten les dades i les metadades en funció de les circumstàncies. Un prototipus ha estat elaborat com una aplicació en l'entorn SAS, per validar conceptes del model.

En el segon projecte sobre metainformació estadística, *EISI (Expert Interface to Statistical Information)* (De Vaney 1992), s'ha elaborat un prototipus, l'arquitectura del qual consisteix en la Interfície d'Usuari i la Base de Coneixement. La interfície és doble: la primera (*Domain Browser*) proporciona mecanismes de navegació i consulta de la base de coneixement, utilitzant un motor hipertext; la segona (*Domain Assistant*) suporta l'accés directe a la metainformació a través de raonaments en el context dels problemes específics del domini d'aplicació, i també mitjançant l'expansió de les especificacions de solucions parcials als problemes, en termes de la metainformació disponible. Pel que fa a la Base de Coneixement, es tracta d'una Base estructurada en tres components: *The usage scenario library, The metainformation database* i *The domain encyclopædia.* La base de dades de metainformació gestiona representacions dels elements, estructures i relacions de metainformació del domini d'aplicació. L'arquitectura d'aquesta base de dades reposa sobre el model elaborat al si del grup de treball METIS: *User's Guide to metaInformation Systems in Statistical Offices.* (UN/ECE 1990).

Un altre dels projectes seleccionats, *CASIP (Complete Automated System for Information Processing)*, descrit per Saris (1992), consisteix en realitat en l'elaboració d'un conjunt de sistemes experts concebuts per suportar cadascuna de les fases de treball d'un camp concret: les enquestes de pressupostos familiars: *EDC (Expert System in Data Collection), MDBS (Micro Data Database System), EDA (Expert System for Data Analysis)* i *EPSS (Expert System for Presentation of Summary Statistics).*

EDA (Expert System for Data Analysis), descrit per Gibert (1992) i Aluja (1993) i elaborat al si del Departament d'Estadística i Investigació Operativa de la UPC, és un entorn integrat per a la producció i l'anàlisi de taules estadístiques i està composat així per dos mòduls: *STG (Statistical Table Generator)* i *STA (Statistical Table Analyzer).* El primer, *STG*, està preparat per a la producció de taules complexes de síntesi a partir de les dades i metadades emmagatzemades pel sistema precedent (*MIDAS*) o bé directament a partir de fitxers ASCII. Al seu torn les taules generades poden ser emmagatzemades a la base de dades estadística del sistema (que inclou també les metadades oportunes) o bé exportades en formats ASCII o LOTUS 1-2-3. El segon mòdul, *STA*, ofereix un conjunt d'eines generals per emmagatzemar, organitzar, tractar, i analitzar gràficament la informació continguda a la base de dades estadística. Els resultats

(dades, gràfics, noves taules...) poden així mateix ser exportats a processadors de textos corrents o a sistemes d'autoedició per a la producció (i eventual publicació) d'informes estadístics.

Com es veu per aquest petit resum, l'activitat en aquest camp és notable. Resta per veure la convergència en la pràctica dels conceptes i funcionalitats involucrats en el desenvolupament de sistemes experts amb els que s'utilitzen en els altres diversos enfocaments, tampoc homogenis entre sí.

2.3. Conclusions de l'experiència internacional

Efectivament, hem pogut constatar la diversitat de plantejaments dels diferents instituts d'estadística, que provenen, tant de la diversitat de situacions de partida com (sobretot) del fet de plantejar-se diversament la problemàtica de la metainformació estadística: bé com un conjunt de problemes particulars o bé com un problema global. Els instituts que prenen aquesta darrera opció resulten ser també els que són més conscients de l'evolució en curs cap a una obertura dràstica dels seus actius d'informació a uns usuaris cada vegada més diversificats i il·lustrats, i de la necessitat de dissenyar i implementar serveis més sofisticats i integrables amb els entorns de consulta i treball dels usuaris. Al mateix temps, són també els més confiats en què la revolució en curs de les tecnologies de la informació i la comunicació i la disponibilitat de sistemes cada vegada més sofisticats fan concebibles solucions pràctiques a aquells objectius.

Ara bé, en la mesura en què s'intenta un plantejament global de la metainformació estadística es manifesta clarament la necessitat d'un model conceptual potent, en el que les diferents tasques, funcions i peces d'informació d'un institut d'estadística tenen el seu lloc, i en el que les múltiples, variades i complexes necessitats dels diferents usuaris estan previstes de manera pràctica i eficient. I això en un context de creixent interdependència i coordinació a diferents nivells, entre instituts, empreses i administracions.

D'aquí la importància de la tasca de reflexió endegada de fa vint anys ençà pels pioners de la metainformació estadística, Bo Sundgren al capdavant.

La conclusió comuna, integrada explícitament per Sundgren als seus treballs, i expressada col·lectivament pels participants al grup de treball METIS, és d'una lògica aclaparadora.

En poques paraules, és la següent:

> *Només si comprenem en totes les seves virtualitats les interrelacions entre les diferents tasques d'informació d'un institut d'estadística i som capaços*

de la seva conceptualització per la via d'un esquema globalitzador, podrem interpretar correctament les interrelacions entre la metainformació estadística i la informació estadística i, per tant, podrem dissenyar sistemes de metainformació estadística integrats.

És clar que si recordem que la necessitat de sistemes de metainformació estadística vé determinada per objectius d'eficiència en l'articulació de sistemes d'informació estadística, ens pot semblar que estem en un cercle viciós si ara diem que el disseny correcte de sistemes de metainformació estadística depèn de la correcta comprensió dels sistemes d'informació estadística en un institut d'estadística. De fet, com fa notar Sundgren (que parla de "simbiosi entre les dades i les metadades"), no ha d'estranyar l'íntima interrelació entre ambdues menes de sistemes perquè és la que existeix entre un sistema d'informació i el sistema objecte sobre el que aquell pretén informar.

Aquesta necessitat no exclou la dificultat de l'empresa, que al seu torn explica les dificultats i vacil·lacions amb què actuen els instituts d'estadística.

Per acabar aquest apartat, ens agradaria fer-ho resumint l'essencial de les observacions que fa Sundgren (1994b) en les seves *Guidelines*[19]. Són les següents:

Molts sistemes de metainformació estadística han fracassat en el passat degut a que:

a) La recollida i manteniment de metainformació estadística és una feina avorrida, cara i llarga en el temps.

b) Desafortunadament, existeix una desconnexió entre els usuaris i els productors de metainformació estadística al si d'un institut. Els que necessiten la metainformació estadística no poden produir-la ells mateixos. D'altra banda, els que "posseeixen" els coneixements sobre les dades estadístiques no troben que en puguin treure massa avantatges de sistemes formalitzats i automatitzats de metainformació estadística.

Així, doncs, des d'una perspectiva positiva, per evitar les males experiències del passat, Sundgren al·ludeix a les següents característiques desitjables del futur SME (Sistema de Metainformació Estadística):

a) Les activitats de recollida de metainformació estadística haurien de ser minimitzades en el sentit de que cap peça de metainformació estadística no hauria de ser entrada més d'una vegada, i les peces susceptibles de ser deduïdes ho haurien de ser de manera informatitzada, i no manualment.

[19] A l'apartat 5.1 (*Experiències del passat i implicacions pel futur*). Aquest resum no pretén substituir el text original, més detallat.

b) Les activitats de recollida retrospectiva massiva de metainformació estadística haurien de ser evitades. En canvi, el major nombre possible de peces de metainformació estadística haurien de ser generades com a subproductes d'altres activitats.

c) S'hauria d'introduir alguna mena d'anàlisi cost/benefici en l'arquitectura d'un SME amb l'objectiu de relacionar usuaris i productors de metainformació estadística d'una manera constructiva.

En definitiva, acaba dient:

> "*Els sistemes d'informació estadística són actius valuosos. Ara bé, sense sistemes de metainformació estadística adequadament integrats, la vàlua d'aquells sistemes es redueix dràsticament. Donat que avui dia els sistemes d'informació estadística estan d'una manera general formalitzats i informatitzats, els sistemes de metainformació estadística han de ser també formalitzats i informatitzats, si es vol que els fluxos de metainformació segueixin el pas de la seva informació objecte, la informació estadística.*"

2.4. El grup de treball METIS (UN/ECE)

2.4.1. Programa de treball

El grup de treball METIS, organitzat al si de la *Conferència d'Estadístics Europeus* (al seu torn dintre de la *Comisió Econòmica per a Europa de les Nacions Unides*), està treballant amb força d'uns anys ençà, per tal de consensuar entre els tècnics representants dels instituts d'estadística dels països europeus (més enllà de la Unió Europea, i als que s'afegeixen regularment a més els de Canadà i dels Estats Units), una sèrie d'aspectes trascendentals sobre la metainformació estadística i el desenvolupament de sistemes de metainformació estadística.

El seu programa de treball, a més de mantenir-se al corrent del d'altres institucions internacionals (en particular, el programa DOSES i, a partir d'ara, el nou programa DOSIS d'EUROSTAT), s'ha concretat especialment en les línies següents:

a) Intercanvi d'informació entre els instituts sobre les línies en les que es basen per a desenvolupar sistemes, globals o específics, de metainformació estadística, i els ensenyaments de les seves respectives experiències.
b) Elaboració d'una terminologia específica comuna per a la metainformació estadística.
c) Inventari i eventual elaboració d'estàndars per ser proposats a les instàncies pertinents.

d) Metodologia per al disseny i desenvolupament de sistemes de metainformació estadística.

e) Criteris d'avaluació de sistemes de metainformació estadística.

f) Mètodes formals per a la descripció de les peces de metainformació estadística.

g) Problemes específics dels països de l'Europa de l'Est per a la transició dels seus sistemes estadístics.

La darrera reunió de treball va tenir lloc a Ginebra, al novembre 1994, mentre la propera no tindrà lloc fins al 1996/97.

2.4.2. METIS i la terminologia

Es tracta d'establir un lèxic de termes relacionats amb la metainformació estadística, i també les seves definicions, com a eina comuna de treball. La darrera versió de l'esborrany, preparada per l'eslovac Prazenka (1994) fou distribuïda a la reunió de novembre 1994, i serà actualitzada i esmenada en el transcurs de 1995.

2.4.3. METIS i els estàndars

Una línia de treball molt important del grup METIS és la relativa a l'estandardització de diversos aspectes de la metainformació estadística.

Una primera tasca consistí en l'elaboració d'un primer element de referència, ja publicat per UN/ECE (1990): *User's guide to metainformation systems in statistical offices*.

Després es va definir la relació entre la metainformació estadística i la informació estadística, al si d'un model de referència general, el *Statistical Reference Model*, proposat per Olenski (1991) en el que es varen identificar tres capes:

- capa de la metainformació estadística (*metadata layer*)
- capa de la informació estadística (*data layer*)
- capa del processament (*computing layer*)

Per a la capa del processament, el punt de partida és considerar un sistema de metainformació estadística, bé com un *Information Resource Dictionary System (IRDS)*, bé com un *Information Resources Management System (IRMS)*, de manera complementària, estàndars en curs de desenvolupament per part de l'ISO[20].

[20] S'està treballant en un estàndar més general, el *PCTE = Portable Common Tool Environment*, del que l'IRDS seria un sub-conjunt, però passaran anys abans que se'n pugui disposar.

D'altra banda, reconeixent que les altres dues capes estan molt relacionades i són ambdues importants per a la metainformació estadística, METIS està treballant en la constitució d'un *Inventory on international standards applicable in statistics* (1a versió, 1993) i en l'anàlisi de l'enfocament adequat per a cadascuna d'elles. Així, l'enfocament basat en el *missatge* s'ha considerat el més idoni per a l'estandardització de la capa de la informació estadística i, per tant, s'ha reconegut la importància de la tasca que ha portat a terme el grup MD6 al si del UN/EDIFACT (Western European Board), que ha elaborat l'estàndar *GESMES (Generic Statistical Message)* per a la definició del missatge estadístic en l'intercanvi electrònic de dades (EDI). Tècnicament, GESMES és un missatge autodefinit, composat de tres parts, de les que les dues primeres són menes de metainformació (administrativa i tècnica, respectivament) referides a les dades, contingudes en la tercera part[21].

En canvi, per a la capa de metainformació estadística, s'ha considerat adequat l'enfocament basat en l'*indicador estadístic* i, conseqüentment, Olenski va elaborar al 1993 una proposta de "model genèric formal per a l'estandardització de la metainformació i els indicadors estadístics", reelaborada i aprovada finalment al 1994. Aquest model, basat en un treball previ en el que es definia un *model semàntic estructural d'indicador estadístic*, accepta l'estàndar GESMES i l'introdueix com un subconjunt. Descriu el seu contingut dient que cobreix els principis de representació de conjunts (holdings) específics de metainformació estadística i defineix regles generals per a l'organització dels indicadors segons diferents formes de missatges: qüestionaris, taules, arxius, gràfiques, etc.

2.4.4. Metodologia per al disseny de sistemes de metainformació estadística

L'objectiu d'aquesta línia de treball ha variat en el temps. Si l'any 1991 s'encarregava a Sundgren que definís "un sistema pilot de metainformació estadística", amb la intenció d'implementar-lo com a banc de proves, al 1993, sota la proposta del mateix Sundgren, l'objectiu es concentrava en els aspectes més genèrics de definir un model de sistema i d'establir unes recomanacions per al disseny de sistemes de metainformació estadística, abandonant justament tota pretensió d'incidir en els aspectes informàtics. A la reunió de novembre 1994, Sundgren (1994) presentava el text definitiu de les recomanacions, sota el títol: *Guidelines for the Modelling of Statistical Data and Metadata*[22], que era aprovat oficialment.

[21] Vegeu-ne una descripció per Maurer (1992). Durant el 1995 s'està procedint a la difusió de GESMES, per a documents (com "*Message Implementation Guide for GESMES/ECOSER*") i programari d'implementació (com "*Windows Help File for GESMES*".

[22] Aquest document és reproduït íntegrament en aquest mateix número de **Qüestiió**.

El raonament de Sundgren, explicitat en diverses ocasions i sobre el que reposen les seves *Guidelines*, comença preguntant-se quin pot ser el concepte de "sistema de metainformació estadística", en el sentit fort de "sistema", més enllà de simples conjunts de metadades, més o menys interrelacionades. El problema consisteix així en trobar els criteris més adequats i eficients que ens permetin articular relacions conceptualment significatives entre les diferents unitats i components d'un sistema de metainformació estadística.

Ara bé, tot sistema d'informació, per definició, es refereix i s'articula entorn d'un sistema objecte, que és el que el justifica i en determina les característiques. Així, *un sistema de* **metainformació** *estadística seria un sistema d'informació l'objecte del qual és un sistema d'***informació** *estadística.* I en la mesura que existeixen diferents menes de sistemes d'informació estadística, també hi haurà diferents menes de sistemes de metainformació estadística.

Per exemple, tradicionalment, els sistemes d'informació dels instituts d'estadística s'han organitzat en general amb una *orientació a la producció (input-oriented)*; és a dir, s'articulen entorn de les operacions de producció estadística i de la sèrie de processos associats a les seves fases.

Ara bé, des del punt de vista de l'ús i dels usuaris de la informació estadística, com subratlla Sundgren, seria "*més adequat organitzar els sistemes d'informació estadística com a sistemes de recuperació i difusió, sobre la base de les necessitats potencials d'informació dels usuaris*". Un sistema d'informació estadística així *orientat a l'usuari (output-oriented)* estaria en les millors condicions per presentar als seus usuaris (tant els tècnics interns com externs) una imatge global ben integrada i conceptualment coherent.

Justament, *organitzar conceptualment i sistemàtica aquesta imatge d'un sistema d'informació estadística és una precondició del sistema de metainformació estadística corresponent, esdevenint així un dels seus objectius inesquivables.*

Així, doncs, per la seva pròpia natura, un sistema de metainformació estadística es presenta al seus usuaris potencials amb la pretensió d'ajudar-los a fer-se una representació mental adequada del sistema d'informació estadística, que és el seu sistema objecte. El resultat esperat és que les accions dels usuaris respecte d'aquest darrer sistema seran així més ajustades, comprehensives i eficaces.

El mètode seguit per Sundgren, en l'elaboració de les seves *Guidelines*, consistirà, doncs, en primer lloc, en definir les *funcions* d'un sistema d'informació estadística i dels diferents *agents* que realitzen les tasques associades en aquelles, en la mesura en què són considerats els usuaris potencials "nats" d'un sistema de metainformació estadística, a afegir als usuaris externs; en segon lloc, tractarà de passar revista a les *necessitats d'informació* (usos) de totes aquestes menes d'*usuaris*, materialitzades,

d'una banda, en certs conjunts o paquets de *peces d'informació* i, de l'altra, en certes funcionalitats o *requeriments del sistema*, que haurien de fer possibles els usos considerats necessaris.

Cal subratllar aquí la transcendència metodològica de considerar, com fa Sundgren, que la funció de *difusió* constitueix una de les funcions normals d'un institut d'estadística, d'on es dedueix que les necessitats de metainformació estadística derivades de les tasques relacionades amb aquella funció són necessitats internes normals de l'institut i no, simplement, necessitats marginals o addicionals.

Efectivament, en la mesura que els agents difusors d'informació estadística d'un institut incorporen en els plantejaments dels sistemes de difusió les necessitats reals dels usuaris externs (i faran justament la seva feina en la mesura en què ho aconsegueixin) podrem analitzar i integrar les necessitats d'aquells sistemes en el plantejament global de les necessitats de metainformació estadística a satisfer pel sistema global de metainformació estadística (SME).

Una altra innovació metodològica de Sundgren en la tipificació d'usuaris d'informació estadística (i, en conseqüència, de metainformació estadística), és la de considerar les pròpies eines de programari (*software tools*) com a usuaris especials, la qual cosa permet analitzar i identificar les seves necessitats sistemàtiques per a integrar-les també en el plantejament global del SME.

En definitiva, a través de la rigorosa consideració dels conceptes implicats en cadascuna d'aquestes fases, Sundgren proposa una sèrie de diagrames de fluxos, que representen models diferents i complementaris d'un sistema d'informació estadística, dels que en dedueix una sèrie de classes de peces de metainformació estadística, per a cadascuna de les quals és aleshores factible definir la informació necessària a recollir en forma de plantilles de documentació esquemàtiques.

Defineix, així, les següents plantilles:

- Declaració de qualitat per a la informació estadística (*micro/macrodata*).
- Documentació d'un arxiu individualitzat de base (*observation register*).
- Documentació d'una enquesta i el seu sistema de producció (*statistical survey*).

Defineix finalment un llenguatge formal de descripció d'objectes de metainformació estadística (en tres capes) i una sèrie de diagrames, que sintetitzen els fluxos d'un sistema de metainformació estadística i n'organitzen l'arquitectura.

Abans, però, fa una sèrie de consideracions sobre les lliçons a treure de les experiències i fracassos del passat, que hem cregut oportú de reprendre a l'apartat 2.3).

2.4.5. Criteris d'avaluació de sistemes de metainformació estadística

Considerant important, malgrat la manca de materials metodològics suficients, l'establiment d'una panòplia de criteris que permetin avaluar l'eficiència i idoneïtat dels sistemes de metainformació estadística, ha estat elaborat un primer esborrany de proposta per De Vaney (1994a) (de World Systems), que defineix l'abast de la tasca i les fases d'avaluació següents: determinació dels requeriments, establiment dels criteris, desenvolupament del model de cost/benefici, eficàcia del procés d'avaluació i anàlisi dels resultats.

Un aspecte polèmic és el de la mesura en què s'han d'aplicar criteris de confidencialitat a la metainformació estadística, en relació amb la informació estadística associada. Hi ha, però, consens en el principi de que "tota la informació i la metainformació estadístiques hauria de ser, en principi, pública; les restriccions només haurien de ser aplicades quan es pugui deduir informació sobre persones individuals".

2.4.6. Mètodes formals de descripció de metainformació estadística

Christofer De Vaney (1994b) ha defensat l'aplicació de les tècniques *FM (Formal Methods)* al desenvolupament de components de metainformació estadística per al seu ús en el context de sistemes de metainformació estadística. El seu raonament és el següent:

En primer lloc, els objectes (o classes d'objectes) de metainformació estadística és probable que siguin utilitzats en la implementació de múltiples sistemes, en diferents entorns. L'ús de tècniques FM en la fase de disseny pot permetre garantir que la semàntica de l'ús és consistent a través de les implementacions.

En segon lloc, en la mesura que els objectes de metainformació estadística i les operacions associades són complexes, les notacions FM poden ser utilitzades com a base per a noves notacions específiques de la metainformació estadística.

En tercer lloc, les tècniques FM poden ser utilitzades com a base per a la definició dels estàndars de metainformació estadística i dels sistemes de metainformació estadística.

En la pràctica, les tècniques FM han estat efectivament aplicades en dos projectes específics relacionats amb la metainformació estadística, en el marc del Programa DOSES d'EUROSTAT: el llenguatge *SMILE, (Statistical Metainformation Language Environment)* desenvolupat, amb l'ajuda d'un sistema expert, per a representar objectes de metainformació estadística i mantenir una base de dades d'aquests objectes, i l'*EISI Domain Assistant*, (*EISI = Expert Interface to Statistical Information Systems*), una extensió de SMILE en el context d'una shell d'inferència per resoldre problemes estadístics (*Vegeu l'apartat 2.2.7*).

En el mateix informe citat, De Vaney donava notícia d'un nou projecte que reprenia el llenguatge SMILE, però basat en els objectes descrits al "Model genèric d'indicadors estadístics" d'Olenski. (1994)

2.5. La metainformació estadística i EUROSTAT

El grup de treball *Metadata Task Force* és tributari, en la definició de la seva feina, de l'activitat estadística d'EUROSTAT que, d'una banda, es concentra en macrodades i, de l'altra, en taules comparables al nivell dels països de la Unió Europea. Així, doncs, els problemes de coordinació i integració d'informació heterogènia, típics de la metainformació estadística, s'agreugen en aquest cas, on l'harmonització de definicions i nomenclatures es troba en el primer pla de les preocupacions, complicades per la necessitat del multilingüisme. (Byfuglien 1994)

L'organització de l'encontre *Statistical Metainformation Systems Workshop*, al 1993, per part d'EUROSTAT (1993), va representar una magnífica ocasió de debatre els problemes de la metainformació estadística en un context obert i tècnic. Ja hem esmentat també (apartat 2.2.7) el programa DOSES de sistemes experts aplicats a la informació estadística.

Tanmateix, als efectes institucionals a escala europea, és el projecte *DSIS (Distributed Statistical Information Systems)* d'EUROSTAT (1992) el que pot tenir efectes importants, en la mesura en que es proposa establir un sistema d'estadística distribuïda entre els països de la Unió. En l'estudi de viabilitat del *DSIS*, es defineix l'anomenat *common reference environment* com a base tècnica que garantiria l'accés a les diferents informacions nacionals sota uns conceptes comuns. Als nostres efectes, cal assenyalar que el *DSIS* reconeix la importància decisiva de la metainformació estadística al si d'aquest entorn comú, a la que fa jugar el paper d'esquema conceptual global, bàsic en la interfície d'usuari del futur sistema. El projecte no ha passat encara de la fase d'estudi, però en tot cas, aquest reconeixement de la metainformació estadística és significatiu. D'altra banda, ja hem esmentat el projecte *EDINOMEN*, independent però integrable amb DSIS, l'objectiu del qual és d'establir un sistema d'intercanvi electrònic de Classificacions i Nomenclatures. (Lebaube 1993)

3. OBSERVACIONS FINALS

3.1. L'estratègia de Sundgren

En una primera versió de les "*Guidelines*", Sundgren (1993b) desenvolupava una sèrie de criteris que configuren una estratègia recomanable per als instituts d'estadís-

tica per al desenvolupament d'un sistema de metainformació estadística. Tot i que aquest text ha desaparegut a la versió definitiva, hem volgut recuperar-lo pels elements pràctics que conté.

El text comença establint les següents premisses:

a) És irrealista pretendre l'immediat desenvolupament d'un sistema de metainformació estadística "complet" que satisfagui totes les necessitats relatives a la metainformació estadística, manifestades en les seves diferents formes. Això sembla evident. Però és que, a més, aquesta afirmació vé confirmada per certes dissortades experiències sofertes pels "impacients".

b) D'altra banda, és igualment arriscat planejar el desenvolupament d'un sistema de metainformació estadística, focalitzat sobre algunes necessitats importants, però sense parar atenció en les exigències i prerequisits relacionats amb elles. Per exemple, "mentre és natural i altament recomanable per a un institut d'estadística modern atorgar prioritat a les necessitats de metainformació estadística dels usuaris d'estadística, fóra estúpid negligir les interdependències entre la tasca subjacent i d'altres funcions d'infrastructura d'un sistema de metainformació estadística més complet". I Sundgren acaba comparant irònicament aquesta manera d'actuar amb la d'aquell que "critica les despeses realitzades en la producció rigorosa d'estadístiques d'atur quan tothom pot llegir les xifres d'atur als diaris cada dia".

Així, la hipòtesi de partida per a Sundgren és que

> "*Cal donar prioritat a les necessitats dels usuaris d'estadístiques, tot i tenint en compte que la metainformació que els usuaris necessiten ha de ser produïda amb criteris econòmics, i també de manera tal que la metainformació i les funcionalitats a obtenir hauran de ser de qualitat*"

En aquest context, cal considerar usuaris tant els interns de l'institut com els externs, i respecte de la metainformació estadística, cal incloure tant les necessitats del seu ús instrumental al si de les diferents funcions (producció, difusió) com les d'ús final dels usuaris externs.

Sundgren preconitza que un institut d'estadística que es proposi desenvolupar un sistema de metainformació estadística segueixi unes fases amb els objectius següents:

1) Definir l'arquitectura global horitzó per al conjunt dels sistemes de metainformació, tal com haurien de ser a llarg termini: el que en podem dir *infrastructura de metainformació.*

2) Implementar aquesta infrastructura de metainformació de manera progressiva, pas a pas, i començant amb aquells subsistemes que:

a) es necessitin amb més urgència per part dels usuaris, o bé

b) es necessitin per a un funcionament eficient dels subsistemes a).

Amb aquests criteris, respecte de les fases del cicle de vida dels sistemes d'informació estadística (d'aquells que es considerin prioritaris), la tasca s'haurà de centrar naturalment en la fase operativa (tant en la perspectiva d'ús com de producció), deixant de banda les fases de disseny i de gestió.

Òbviament, d'acord amb el que hem dit suara, la prioritat aniria als subsistemes més enfocats als usuaris i, en segon lloc, a aquells subsistemes orientats a la producció, però que siguin essencials per al bon funcionament d'aquells.

3.2. Conclusions

Sigui quina sigui la solució adoptada per a la implementació d'un SME, que no podrà ser d'altra banda sinó progressiva, ens permetem subratllar les següents condicions, que entenem com a necessàries, si bé no suficients, per garantir l'eficiència i el reeiximent de l'operació.

1) Que el disseny del SME a llarg termini sigui global, respongui a un esquema conceptual rigorós i es defineixin les interrelacions entre els subsistemes existents, o a crear progressivament, i també entre les diferents peces de metainformació estadística, a la llum d'aquell model global, que podem considerar, d'altra banda, simplement com un sistema virtual, que pot no arribar a implementar-se mai.

2) Que en la selecció, disseny i implementació dels subsistemes reals, components del SME, l'orientació a l'usuari sigui prioritària, tot i garantint, en la mesura necessària, l'eficàcia de la producció d'informació estadística.

3) L'esmentada orientació a l'usuari hauria de partir de la consideració a llarg termini del conjunt de les necessitats del Sistema Estadístic del país o territori de l'institut d'estadística, independentment de l'abast de les seves competències. Com a conseqüència, les necessitats de coordinació en cadascuna de les fases de les activitats dels diferents organismes productors d'estadística oficial del país formaran part del SME global.

4) Donades aquestes característiques del SME global, en el seu disseny hauríen de participar activament els tècnics més propers i sensibles a les necessitats de l'ús de la informació i la metainformació estadístiques en les diferents fases i funcions de l'institut. D'altra banda, la responsabilitat del disseny hauria d'estar situada orgànicament sota la màxima autoritat de direcció.

5) Donades les incerteses de partida, pot ser prudent triar els components a implementar primer amb el criteri de relativa autonomia respecte dels sistemes

centrals, per tal de no interferir amb les tasques bàsiques d'un institut d'estadística. Això pot portar a experimentar sistemes d'edició electrònica, bé sigui en CD-ROM, bé en interfícies gràfiques d'usuari a bases de dades (eventualment en entorn WWW).

6) Paral·lelament, s'haurien d'implementar els subsistemes relacionats amb l'eficiència conjunta de la producció, sobretot els necessaris per a la coordinació tècnica interna. Caldrà fer-ho, però, dintre de l'esquema global i conceptual de referència, per tal de garantir la coherència lògica entre els diversos components.

7) Finalment, en l'estat actual de les coses, en el que el tema de la integració de la metainformació estadística en els sistemes d'informació estadística està a les beceroles, caldrà fer un seguiment atent dels desenvolupaments que es produeixin a nivell internacional, tant en el terreny dels estàndars com en el de les aplicacions pràctiques. És a dir, fer el que se'n diu una "vetlla tecnològica" atenta sobre el tema.

REFERÈNCIES

[1] **Aluja T, Balaguer J.M., Martí-Recober M., Nafria E.** (1993). "The EDA/System. A system for the production and analysis of summary statistics". *Statistical Metainformation Systems Workshop. Proceedings.* (EUROSTAT) Luxembourg, February 2–4, 1993.

[2] **Appel, Günther** (1993). (Statistiches Landesamt Berlin). "A metada driven statistical information system". *Statistical Metainformation Systems Workshop. Proceedings.*

[3] **Appel, Günther** (1994). (Statistiches Landesamt Berlin). "Management aspects of a statistical metainformation system (SMS)". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 6. Geneva : UN/ECE, 1994.

[4] **Byfuglien, Jan** (1994). (Eurostat). "Some plans and issues developing metadata at Eurostat". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 12 Geneva : UN/ECE, 1994.

[5] **Canals, Isidre** (1992). (Institut d'Estadística de Catalunya). "Los sistemas hipertexto e hipermedia en el contexto de los futuros libros electrónicos. Reflexión conceptual ilustrada con casos prácticos". *4as Jornadas de Información y Documentación en Ciencias de la Salud.* Bilbao, Junio 1992.

[6] **Darius P., Boucneau M., De Greef P., De Feber E., Froeschl K.** "Modelling metadata". *Statistical Metainformation Systems Workshop. Proceedings.*

[7] **De Vaney, Christofer** (1994a). (World Systems). "Evaluation criteria of statistical metainformation systems". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 20. Geneva : UN/ECE, 1994.

[8] **De Vaney, Christofer** (1994b). (World Systems). "Formal Methods and the design of statistical metadata". *UN/ECE. METIS Meeting*. Nov. 94. Working Paper 22. Geneva: UN/ECE, 1994.

[9] **De Vaney, Christofer** (1994c). (World Systems). "EMMA. Enhanced metainformation Management Architectura". *UN/ECE. METIS Meeting*. Nov. 94. Working Paper 14. Geneva: UN/ECE, 1994.

[10] **Dippo, Cathryn S.** (1994) (US Bureau of Labor Statistics). "A customer focus on metadata". *UN/ECE. METIS Meeting*. Nov. 94. Working Paper 11. Geneva: UN/ECE, 1994.

[11] **D'Angiolini, Giovanna** (1993). "Data and metadata: The need for a knowledge approach". *Statistical Metainformation Systems Workshop. Proceedings.*

[12] **EUROSTAT** (1992). *DSIS — Distributed Statistical Information System. Feasibility Study*. Version 2. Luxembourg : Logica, 1992 77p. Annexos.

[13] **EUROSTAT** (1993). *Statistical Metainformation Systems Workshop. Proceedings*. Luxembourg, February 2–4, 1993.

[14] **Gibert C., Martí-Recober M., Aluja T.** (1992). *EDA: An Expert System for Data Analysis*. Barcelona : UPC, Dept. d'Estadística i Investigació Operativa, Maig 1992. Document de recerca.

[15] **Gillman, Daniel W. & Appel, Martin V.** (1994). (U.S. Bureau of the Census). "Metadata Database Development at the Census Bureau". *UN/ECE. METIS Meeting*. Nov. 94. Working Paper 10. Geneva : UN/ECE, 1994.

[16] **Graves, Ronald** (1994). (Statistics Canada). "Information holdings within Statistics Canada. A framework". *UN/ECE. METIS Meeting*. Nov. 94. Working Paper 18. Geneva: UN/ECE, 1994.

[17] **Hand, D.J.** (1992). "Artificial Intelligence in statistics". *New Technologies and Techniques for Statistics. Proceedings*. EUROSTAT. Bonn, 24-26 Feb. 1992. p. 133–140.

[18] **INE** (1994). *S.I.D. Manual de referencia y uso. Documentación del subsistema de gestión*. Madrid : Instituto Nacional de estadística, 1994. (Doc. interno).

[19] **INSEE** (1990). *Les Dictionnaires de Données Statistiques (DDS). Présentation et Manuel de Référence*. Paris: INSEE, 1990

[20] **Kopp, Norbert** (1993). (Statistiches Landesamt Berlin). "The Thesaurus as a user interface for the metainformation system". *Statistical Metainformation Systems Workshop. Proceedings.*

[21] **Lamb, Joanne** (1993). "Metada in survey processing". *Statistical Metainformation Systems Workshop. Proceedings.*

[22] **Lazarou, Georges** (1993). (INSEE). "Les Dictionnaires de Données Statistiques (DDS)". *UN/ECE. METIS Meeting*. Dec. 1993. CRP 8. Geneva: UN/ECE, 1993.

[23] **Lebaube, Philippe** (1993). (EUROSTAT). "Exchange systems for classifications and standardization issues". *Statistical Metainformation Systems Workshop. Proceedings.*

[24] **Malmborg Eric** (1988). (Statistics Sweden). "Design of the user interface for an object- oriented statistical database". *Statistical and Scientific Database Management, Fourth International Working Conference SSDBM*, Rome, June1988. Available: Statistics Sweden R & D Report 1988:11. (cit. Malmborg 1993).

[25] **Malmborg E. & Lisagor L.** (1993). (Statistics Sweden). "Implementing a statistical metainformation system". *Statistical Metainformation Systems Workshop. Proceedings.*

[26] **Marina, Liana** (1994). "Using Statistical Data Dictionary (DDS) client/server towards a Statistical Metainformation System in NCS, Romania". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 23. Geneva: UN/ECE, 1994.

[27] **Maurer A., Kubler J.E.** (1992). "EDI and statistics. The need for a Generic Statistical Message". *New Technologies and Techniques for Statistics. Proceedings.* EUROSTAT. Bonn, 24-26 Feb. 1992. p. 258–267.

[28] **Nordbäck, Lars** (1990). (Statistics Sweden). *An illustrated booklet on the dissemination of statistics in the 1990s.* Stockholm : Statistics Sweden, 1990.

[29] **Nordbäck, Lars** (1992). (Statistics Sweden). "The PC-AXIS vision, the liberation of official statistics". *New Technologies and Techniques for Statistics. Proceedings.* EUROSTAT. Bonn, 24-26 Feb. 1992. p. 218–225.

[30] **Nordbotten, Svein** (1993). (Univ. Bergen). "Statistical meta-knowledge and -data". *Statistical Metainformation Systems Workshop. Proceedings.* Luxembourg, February 2-4, 1993.

[31] **Olenski, Jozef** (1991). (CSO Poland). "Reference Model for Standardization in Statistics". *UN/ECE. METIS Meeting.* Dec. 91 Working Paper 3. Geneva: UN/ECE, 1991.

[32] **Olenski Jozef** (1994). (CSO Poland). "Standardization of Statistical Indicators and Metadata ". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 17 Geneva: UN/ECE, 1994.

[33] **Prazenka Dusan** (1994). "Common Terminology of METIS". (2nd version). *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 16. Geneva : UN/ECE, 1994.

[34] **Sadreddini M.A., Bell D.A.** (1993). "Integrating distributed and heterogeneous statistical data bases". *Statistical Metainformation Systems Workshop. Proceedings.*

[35] **Saris Willem E., Prastacos Moulicos, Martí Recober Manuel** (1992). "CASIP - A complete automated system for information Processing in family budget research". *New Technologies and Techniques for Statistics. Proceedings.* EUROSTAT. Bonn, 24-26 Feb. 1992. p. 80–87.

[36] **Schuerhoff M. & Bethlehem J.G.** (1993) (Netherlands Central Bureau of Statistics). "Handling data and metadata in BLAISE III". *Statistical Metainformation Systems Workshop. Proceedings.*

[37] **Shoshani A., Wong H.K.T.** (1985). "Statistical and scientific database issues". *IEEE Transactions on Software Engineering*, SE 11 (10), 1985. (cit. D'Angiolini).

[38] **Silver, Mick** (1993) (Cardiff Business School). "The role of footnotes in a statistical metainformation system". *Statistical Metainformation Systems Workshop. Proceedings.*

[39] **Sundgren, Bo** (1973). (Statistics Sweden). "An infological approach to data bases". *Urval 7* (cit. Nordbotten).

[40] **Sundgren, Bo** (1990). (Statistics Sweden). *Conceptual modelling and related methods and tools for computed-aided design of information systems* Stockholm : SCB. R & D Report, 1990.

[41] **Sundgren, Bo** (1991a). (Statistics Sweden). *What metainformation should accompany statistical macrodata?* Stockholm : SCB. R & D Report, 1991:9.

[42] **Sundgren, Bo** (1991b). (Statistics Sweden). *Statistical metainformation and metainformation systems.* Stockholm : SCB. R & D Report, 1991:11.

[43] **Sundgren, Bo** (1992a). (Statistics Sweden). *Organizing the metainformation systems of a statistical office.* Stockholm : SCB. R & D Report, 1992:10.

[44] **Sundgren, Bo** (1992b). (Statistics Sweden). *Some properties of statistical information: Pragmatics, Semantics and Syntactics.* Stockholm: SCB. R & D Report, 1992:16.

[45] **Sundgren, Bo** (1993a). (Statistics Sweden). "Modelling Metainformation Systems". *Statistical Metainformation Systems. Workshop. Proceedings.* Luxembourg, February 2-4, 1993.

[46] **Sundgren, Bo** (1993b). (Statistics Sweden). *Guidelines on the design and implementation of statistical metainformation systems.* Stockholm : SCB. R & D Report, 1993:4.

[47] **Sundgren, Bo** (1994a). (Statistics Sweden). "Statistical metadata and metainformation systems". *UN/ECE. METIS Meeting.* Nov. 94. Working Paper 15. Geneva: UN/ECE, 1994.

[48] **Sundgren, Bo** (1994b). (Statistics Sweden). *Guidelines for the Modelling of Statistical Data and Metadata.* Stockholm : SCB. R & D Reports, 1994. (Reproduït íntegrament en aquest número de **Qüestiió**).

[49] **UN/ECE** (1990). *User's Guide to Metainformation Systems in Statistical Offices. (METIS).* Geneva: UN/ECE, 1990. (cit. De Vaney).

[50] **Villán, I.** (1991) (INE). "Metadata management at the Spanish National Institute of Statistics: SID project". *UN/ECE. METIS Meeting.* Dec. 91 Working Paper 11. Geneva: UN/ECE, 1991.

[51] **Walker D., Newman I., Mather P., Ruggles C, Medyckyj-Scott D.** (1993). "Federal network based metainformation and central metadata management. Some initial proposals related to the GENIE approach". *Statistical Metainformation Systems Workshop. Proceedings.*

[52] **Zeisset, P.T.** (1993). (US Bureau of the census). "Metainformation for summary statistics: The EXTRACT experience". *Statistical Metainformation Systems Workshop. Proceedings.*

ENGLISH SUMMARY:

INTRODUCTION TO STATISTICAL METAINFORMATION SYSTEMS

Isidre Canals Cabiró

This paper aims to provide an introduction to the concept and problems related to statistical metainformation, defined broadly as *information on statistical information.*

The first part ("Introduction to statistical metainformation systems") takes the perspective of an official statistical body faced with the problem of organizing statistical metainformation (metadata) in order to provide its users with all additional information needed to interpret and use statistical information (data). These complementary elements include the methods used for producing that statistical information, and also definitions, classifications, territorial codes etc.

It is stated that statistical metainformation (or metadata) can be analyzed from two points of view: from the user and the producer perspectives, which induces two types of statistical metainformation system: user-driven and production-driven. In fact, both internal and external users should be considered as potential users of a statistical metainformation system.

From the viewpoint of the use of statistical information the paper reviews the new means of disseminating information (electronic books on CD-ROM, online databases...) and raises questions on how to integrate statistical metainformation with the data. It is also stated that the user perspective on metainformation can be fruitfully used to design the statistical information system of a statistical body, thus creating a *user-oriented statistical information system.* The complexity of this task can be recognized if we take into account the fact that potential users are heterogeneous in their needs and that now, more than ever, they possess the skills and the technical means to manage statistical information by themselves.

From the viewpoint of the production of statistical information, this being the basic function of every statistical body, the statistical metainformation of interest to us will be all the additional information necessary to produce statistical information correctly and efficiently. So, all the information needed at every stage of the production chain (design, data collection, primary registers, estimation, tabulation) and also dissemination stage will be considered items of metainformation. Technical documents relating to statistical operations, and information related to different types of tools (computer systems, classifications, codes...) are typical examples of the elements of *production-driven statistical metainformation systems.*

Both viewpoints, user and producer, not only imply different statements of the problem, but they also convey different needs, according to different functions and tasks of the users (both internal and external), who require different items of information about the same subject. Nevertheless, both approaches may converge in the future, as they are related to complementary functions in a statistical body. Moreover, the organization of the information concerned will have a tendency to be conveyed to the users through a consistent global conceptual scheme.

This convergence will also be driven by the fact that users will be both, internal and external. Broadly speaking, external users will be concerned basically with the correct end use of data, while internal users will be concerned with the efficient production of data.

The paper proceeds to analyze and classify the different types of statistical metainformation, according to their nature, format, support and intended goal. The resulting strong heterogeneity of this information constitutes an added difficulty to the problem of organising a statistical metainformation system. A series of examples in different contexts of use of statistical metainformation serves the purpose of illustrating the variety of uses and users of statistical metainformation.

The first part ends with the following definition by METIS (the working group of the Conference of European Statisticians):

> *A Statistical Metainformation System (SMS) is the information system that uses and stores statistical metadata and produces statistical metainformation for purpose of supporting decision making concerning an information system. The object of the Statistical Metainformation System is the statistical information system.*

The second part of the paper ("The diversity of international experience") is devoted to a survey of some projects and experiments carried out by different statistical bodies in the field of organising statistical metainformation systems, and also of proposals made by international bodies.

The projects and experiments reviewed have been grouped, according to the method or approach adopted, as follows:

1) General production-oriented statistical metainformation systems. (The "Dictionnaires de Données Statistiques (DDS)" and information management approaches)

2) Specialized statistical metainformation systems.

3) General user-oriented statistical metainformation systems. (Integrated user interface, PC-DOK system and EMMA, Enhanced Meta- information Management Architecture)

4) Document systems and statistical metainformation systems.

5) Statistical metainformation as associated with sets of statistical information on disseminating digital supports. (Diskettes, CD-ROM, Public interactive databases and Internet World Wide Web servers)

6) Statistical metainformation system as a coordination tool

7) Statistical metainformation, integrated in an expert system

The second part also describes the workings of and the opinions issued by specialised working groups of international bodies. Work programs of METIS are described briefly, especially those relating to the methodology of design and implementation of statistical metainformation systems. The proposal written by Sundgren ("Statistics Sweden"), officially adopted by METIS as guidelines is reproduced in this volume of **Qüestiió**.

The third and last part of this paper repeats some recommendations made by Sundgren to the statistical bodies on their strategy for the development of statistical metainformation systems, and summarizes some general conclusions, emphasizing the need for a long-term global design of the system, even in the (advisable) case of its partial and progressive implementation.