

SEMINAIRE

E+; 3=! C

SEMINAR

STATISTICAL COMMISSION AND
 ECONOMIC COMMISSION FOR EUROPE



Distr.
 GENERAL

CONFERENCE OF EUROPEAN
 STATISTICIANS

CES/SEM.43/6
 4 May 2000

ENGLISH ONLY

Seminar on integrated statistical information
 systems and related matters (ISIS 2000)
 (Riga, Latvia, 29-31 May 2000)

Topic I: Data warehousing and the development and use
 of statistical databases in a network environment

**THE IDEAL WAREHOUSE - INTEGRATED, INTEROPERABLE, INTERDISCIPLINARY
 AND INTERNATIONAL**

Invited paper

Submitted by the Office for National Statistics, United Kingdom¹

I. INTRODUCTION

1. Data Warehousing originally took root in the world of business and commerce, particularly in the areas of resource allocation, stock control and customer responsiveness where, in successful implementations, it has generated enhanced efficiency, improved competitiveness and greater profitability. The concept has since been taken up by Government, and particularly by Government Statistical Organisations (GSOs), where it has been hailed as a more effective way of storing, amalgamating, managing and exploiting the vast amounts of data handled by government statisticians.

2. Within this GSO context a statistical Data Warehouse can be defined as:

A single, complete and corporate repository of linked metadata and data which have been trawled from all of a GSO's disparate data sources, assembled, amalgamated and documented in a standard format, and stored in a structure which allows users to view, query, mix-and-match, and download material for analysis at different levels of access depending on their degree of authorisation.

3. This definition implies that the major role of a Data Warehouse within the statistical production process is to combine data and make it as accessible as possible to as many people as possible. However, as this paper emphasises, a Data Warehouse can also play an important role as a management tool, particularly in the areas of Data Management and Knowledge Management.

1 Prepared by James Denman.

4. Although the views expressed in this paper are the author's own they also reflect much of the strategic thinking which underpins the UK's Office for National Statistics' own response to a range of recent developments which impinge on the world of government, government statistics and informatics. So what are these various 'Drivers'?

II. BACKGROUND

Social, Cultural and Market Factors

5. Within the next few years some 10 million people in Britain are expected to be linked to the Internet and the pace of Web connections in the UK is expected to accelerate with the introduction of unmetered access, the replacement of analog by digital transmission, and the take-up of newer and faster technologies such as Wireless Application Protocol (WAP), Assymetrical Digital Subscriber Lines (ADSL), and other, even faster processes. There are already some half a million websites in Britain, and their numbers are expected to double in the next four years. By 2005, 130 million Europeans are expected to access the Internet regularly via mobile phones and, by 2020, according to Microsoft, nine out of every 10 books sold will be electronic. These far-reaching technological developments are reflected in the British Government's own agenda.

Central Government Policy Imperatives

6. Government thinking in the UK is spearheaded by a top-down drive to link every household and every school to the Net with free access provided to those unable to afford it themselves. This overall vision is reinforced by a legislative drive to release as much government information as possible under the proposed Freedom of Information Act, and complemented by a policy drive to make government itself become more 'On-line'. It is also underpinned by a push for the greater integration of government services ('Joined-Up Government') and the provision of more visible, more user-friendly and more accessible services ('E-Government'). Metadata, which can improve awareness and foster understanding, and turn 'data' into 'information', plays a pivotal role within this scenario because it seen as the main communication bridge between government and the citizen. A drive to provide comprehensive metadata about the whole of the UK government, and all government services forms a key element in this whole process. Within this scenario, every Government Department is expected to meet the following targets:

- 25% of government services to be available electronically by 2002,
- 100% of government services to be available electronically by 2005.

Producer/User Requirements

7. Users of statistics and many producers of statistics are demanding a more 'creative' approach to government statistics, away from the traditional 'push' culture and towards more of a 'pull' culture. Their desire is for government to share its statistical resources more widely, to make them more available and more accessible for cross-cutting analysis by opening them up further down the production line, and further down the aggregation hierarchy.

Emerging Technologies

8. These internal UK initiatives are reinforced by an international drive to encourage greater inter-operability of national information resources using communication technologies and transfer protocols such as XML and Z39.50 and information structures such as the Resource Description Framework (RDF).

III. THE UK STATISTICAL AGENDA

9. The impact of all these interacting factors - the growth of the Web, increasing demand for easy access to integrated information resources, the emergence of standard communication technologies, the widespread adoption of metadata standards and structures, - lies at the top of the statistical agenda within the UK. The Office for National Statistics (ONS) and its sister Statistical Divisions in each of the 30 Government Departments which make up the UK Government Statistical Service (GSS) 'own' a great deal of the raw material which constitutes the currency of what has been termed the 'Information Age'. They also have a crucial part to play in the drive to make government more available and more understandable to the 'governed'. All these developments underpin the ONS/GSS's new 'National Statistics' Code of Practice which will be launched in June of this year and which involves, inter alia:

- a greater emphasis on hallmarking statistical collections and outputs against agreed quality and methodology standards
- accelerated efforts to encourage the harmonisation and integration of hitherto separate statistical resources
- a drive to make statistics more widely available and accessible in a way that facilitates cross-cutting analyses
- efforts to produce a finer geographical disaggregation of statistical outputs in order to make them more 'local' and thus more relevant to citizens' own experience
- a gradual move away from chargeable hardcopy dissemination and towards free, on-line, Web-based distribution of statistics
- the introduction of electronic books - downloadable and free of charge
- moves to embrace e-commerce - e.g. giving customers the opportunity to order hardcopy books or bespoke analyses on-line.

10. The birth of the Information Age has also emphasised the need for more effective Information and Knowledge Management procedures. This has raised the big question - how best to organise and engineer the process of collecting, compiling, processing, analysing, interpreting or disseminating data in order to meet the target of 'Electronic Government' and best suit the needs of data providers, data intermediaries, and end-customers of data? These questions are explored below.

IV. THE GOVERNMENT STATISTICIAN'S PRODUCTION ROLE

The Data Transformation Cycle

11. At the most basic level the role of the government statistician or producer of statistics is to collect raw data and transform that data into coherent information which users can understand and exploit in order to expand the sum total of human knowledge. This seems a straightforward enough process but, as Table 1 shows, the whole data-information cycle involves a host of other activities and interfaces, which can be categorised under the following main functional headings:

Take-On
 Primary Processing
 Standardisation and Packaging
 Secondary Processing
 Aggregation, Expansion, Amalgamation and Analysis
 Dissemination

12. The full range of activities described in Table 1, especially within in a federal statistical organisation like the UK's, tend to be undertaken using a whole variety of different operational systems all of which use a mixture of different database environments. These first-tier data collection systems also tend to be

isolated from one another because they are held in fundamentally incompatible locations and formats and on a variety of platforms. As a result their contents have to be re-assembled or re-engineered within a variety of other operational systems, at both the local, intermediate and central level before the data can be amalgamated and exploited for further analysis or for distribution through a whole range of dissemination media such as publications, on-line databsases, CD-Roms, etc. These second-tier compendium systems create, in their turn, a second tier of disparate and incompatible sources of information because they too use different technical architectures.

13. The challenge facing information managers within government is how best to bring together information which is isolated in a variety of different source or operating systems and integrate that information in a way which best meets users' needs. Data Warehousing has been proposed as the optimum solution.

The Dissemination Role

14. Of all the activities described in Table 1, the one function which seems to be ripe for greater rationalisation is the 'Dissemination function' whether it involve metadata dissemination or data dissemination. Experience shows that the provision of metadata is either, not done at all, or left until the end and performed as an afterthought. More energy and enthusiasm is directed at the dissemination of data, but in recent years even this function has become a burden. This is because the data producer who wants to supply data as inexpensively and as quickly as possible, now finds himself or herself confronted with a much broader and more complex requirement than ever before, having to supply data to a whole range of different media, none of which are necessarily integrated from either a managerial or editorial point of view, e.g.

- Press Releases (for immediate dissemination)
- Fax services (ditto)
- Web pages (ditto)
- On-line Databases (ditto)
- Tailored outputs
- Subject-specific publications (for regular but less immediate dissemination)
- Compendia publications (for periodic dissemination)
- CD-Roms / Diskettes (for periodic dissemination)
- Ad hoc outputs.

15. This state of affairs can be characterised by the 'spaghetti junction' effect shown in Chart A. Here again the hope and expectation is that Data Warehousing can induce the traffic-calming solution shown in Chart B.

V. THE GOVERNMENT STATISTICIAN'S DATA MANAGEMENT ROLE

16. The data transformation process described in Table 1 is normally undertaken in the context of a standard code of practice which obliges government statisticians to work within certain management parameters involving, for instance:

Professionalism - i.e. managing the processes of collection, compilation, storage and dissemination in a way that best meets user requirements for coverage, relevance, accuracy, timeliness, comparability, consistency, coherence, and disaggregation.

Cost Effectiveness - i.e. reconciling the aims of minimising the costs to providers and maximising the benefits to users with a view to achieving the best possible value within the technology and resources available.

17. These traditional roles has been supplemented in recent years by an increasing emphasis on the 'Data Custodian' role. This role is based on a growing recognition

that statistical data, like many other resources, are a valuable commodity, are often produced at considerable expense, can be graded as non-renewable, may be irreplaceable if neglected, damaged or lost, and can only realize their maximum value if properly managed and nurtured, and if exposed to widespread and long-term use. Out of this realization has come a recognition of the need for the producers of statistics to devote more of their energies to the Information Management function, embracing each of the following stewardship responsibilities:-

- Custodianship
- Standardisation and Harmonisation
- Quality Assurance
- Statutory Compliance
- Documentation
- Metadata for Assessment and Evaluation
- Risk Control
- Retention and Preservation
- Audit and Review
- Access and Dissemination Control

18. Ideally the whole data management process described in Table 2 should be undertaken right at the beginning, and at every subsequent stage, of the data transformation cycle rather than left until the end. In practice, however, this is not the case. Furthermore, the degree of importance attached to each of these functions varies from individual to individual, business to business. Many statisticians plead lack of resources and lack of time. Most require help to facilitate the data management function and many would welcome a central and on-line data management system which would allow them, for example, to:

- trawl for relevant information (e.g. a question bank),
- refer to appropriate reference information,
- have on-line access to standard documentation/metadata templates.

VI. THE GOVERNMENT STATISTICIAN'S CUSTOMER SERVICING ROLE

19. A typical Government Statistical Organisation's customers fall into a wide spectrum of categories encompassing colleagues and policy makers within government, Local Authority, Health Authority and other quasi-government officers, the media, financial analysts, economists, businesses, academics and teachers, genealogists, students and the general public outside government. Each varies in the speed with which they require statistics and the depth of detail they require but all customers share certain common requirements. At both the national and international level, all customers seek to a greater or lesser degree :

- an easy means to discover the existence and whereabouts of statistics,
- an easy and single point of on-line access to multiple and linked resources,
- a simple-to-use Interface,
- An indication of the useability and quality of those resources,
- metadata embedded within the data so that they can better understand the statistics,
- machine to machine transferability allowing data to be downloaded in a variety of formats whether as text, tables, or maps so that users can use systems and tools with which they are already familiar,
- a DIY (Do-It-Yourself) interface i.e. a system which allows users to mix and match statistics from different types of source and different countries in order to produce multi-disciplinary cross-analyses,
- data disaggregated to the small-area level or at least to the level of commonly-used geospatial units,
- data which can be obtained in both hardcopy and electronic format - i.e. 'Books

and clicks',

- electronic books which are dynamically updated,
- e-commerce facilities which allow them to purchase chargeable products on-line.

20. This is a fairly tall order and, to their credit, government statisticians have made great strides in recent years in their attempts to meet these needs, by:

- harmonising the concepts and questions used in censuses and surveys to ensure inter-operability and coherent outputs,
- publishing more and more metadata alongside their data,
- placing more material on the Internet in a format which can be downloaded for further manipulation.

21. However many of the products of government statisticians have a long way to go to lend themselves to cross-cutting analysis. This is because they still tend to be:

- input rather than output-focused and developed according to what some have called the 'Silo' or 'Stovepipe' approach to information management,
- producer-focused rather than customer-focused,
- pre-packaged and ready-made rather than DIY-ready,
- Static rather than dynamic.

22. From the users's point of view, the ideal world would be one in which he or she can log into a search tool which will locate any statistical resource anywhere in the world which contains the sort of statistical material they need and transport that material back to their own PC. Ideally they want to be able to access a single repository of metadata and data containing a comprehensive array of official data covering all areas of statistical life and which have been:

- collected using standard definitions,
- collated according to standard dimensions,
- described, defined and documented in standard ways,
- disaggregated to standard geospatial levels,
- stored in standard formats,
- assembled for searches using standardised techniques,
- prepared for transfer/download according to standardised procedures.

23. From the users' point of view the idea of Data Supermarket or Data Warehouse seems to offer the best chance of meeting their needs as well as the needs of the producers themselves.

VII. ACHIEVING THE IDEAL

24. The kind of Warehouse required to match the concept outlined in this paper's opening paragraph, and meet the demands of producers and consumers described in subsequent paragraphs will typically require a three-tier structure combining hardware, software and connectivity products, and containing both metadata and data. Typically it should encompass:

- An **Input** facility - for extracting and importing data from outlying 'Source', 'Operating' or 'Production' systems and transforming it for Warehouse usage.

- A **Data Management** facility - for laundering, organising, integrating, amalgamating, standardising, aggregating and storing multiple data types within a single infrastructure.

- A **Dissemination** facility - to provide users with a range of access routes to the data - incorporating 'mining' tools to enable producers to dredge the repository

in order to design and deliver different types of products, and 'harvesting' tools to allow users to access, query, view and analyses data in a convenient form.

25. The sheer scale and cost of building the sort of system described above, and its likely disruption to day-to-day operations is enough to dissuade most GSOs from making the necessary investment in a full-scale 'Big-Bang' Data Warehousing approach. Some have been tempted by a more incremental, more scaled-down approach based on the idea of 'Data Marts' (or small warehouses) or linked 'Data Sheds' which house smaller, more summary datasets.

VIII. ONS EXPERIENCE WITH DATA WAREHOUSING

26. To date, the UK'S Office for National Statistics has opted for neither of these alternatives but has instead gone down the route of creating a network of separate, independent and 'lateral' warehouses, each of which serves a different purpose and meets a specific need, and none of which matches the exacting specification of full-blown longitudinal warehouse described earlier in this paper. The ONS's current 'warehouse network' is made up of four basic components:

The Central Shared Database (CSDB) - this is an 'operational warehouse' which holds economic data in the form of time-series. The CSDB acts as the cooking-pot for macroeconomic aggregates such as the UK's National Accounts and Balance of Payments estimates, and a source repository for all the (mainly economic) material which appear as an output in Press Releases, hardcopy publications and on-line data delivery channels. The latter comprise:

DataBank: a subscription service which provides for bulk delivery of over 25,000 of the CSDB's time-series within electronic 'parcels' whose contents replicate hardcopy publications.

TimeZone: a 'Son-of-DataBank' service which releases the same content but on a 'mix-and-match' and 'per-series' basis.

The Regional Statistics Database (RSDB) - this is a 'publication warehouse which is still under development and which holds mainly cross-sectional (i.e. multi-dimensional) datasets imported from outlying operational systems. The RSDB has a largely social or socio-economic content and its prime purpose is to store and amalgamate all the 'core' datasets (those used every year) which are required as content for ONS's main Compendia or cross-cutting publications such as 'Social Trends', 'Regional Trends', etc. The RSDB will eventually be extended to cover the complete portfolio of compendia publications and has also been designed to provide content for on-line delivery channels such as StatBase and the ONS Website. The RSDB has gone a long way towards reducing data supply inefficiencies by reducing the need for data contributors to supply the same data several times to separate Publication Editors at different times of the year. It depends on a complementary 'Data Collection Strategy' which now applies to all core datasets which feature in selected corporate outputs. The traditional 'Collect-once / Supply many times' routine (illustrated in Chart A) has now been replaced with a new routine based on the concept of 'Collect-once / Supply once / Use many times' (illustrated in Chart B).

StatBase - this is an 'electronic dissemination warehouse' which holds both metadata and data in a standard structure for on-line access over the Web. StatBase already imports much of its economic time-series content from the CSDB and will, in future, derive much of its social cross-sectional content from the RSDB. The system also allows for metadata and data to be supplied via in-house software supply tools known as the 'Metadata Assistant' and 'Data Assistant'.

27. All three systems serve a number of useful functions and go a long way towards helping the ONS and the GSS to meet many of customer requirements outlined in earlier paragraphs, e.g.:

- Enhancing users understanding of statistics (through metadata provision)
- Improving the availability and speed of access to statistics (through the Web)
- Increasing the inter-operability of statistics (through standard structures)

IX. LESSONS LEARNT

28. However, none of ONS's systems, either individually or in tandem, provide an adequate Input facility or Data Management facility. As a consequence, they fail to deliver the main benefits trumpeted by exponents of the Data Warehousing option. Their major weaknesses are their 'lack of connectivity'.

29. There are either inadequate or non-existent links between the three warehouses themselves, between the warehouses and outlying operational systems, and perhaps most importantly, between the warehouses and ONS's Website. StatBase, for instance, has helped the data user but has provided little direct help to the data provider. On the contrary many providers see it as a burdensome 'add-on' to the producers responsibilities rather than a tool which makes it easy for the producer to fulfill his obligations, 'Unfriendly' supply facilities are the main problem here.

30. Each system offers data producers only limited help with their Information Management function and thereby fails to deliver what could be one of the main attractions of a Warehouse - the ability to offer everything a data producer needs in the way of data management facilities, at the click of a button. Without this kind of buy-in, support has been less than fulsome.

X. A NEW APPROACH

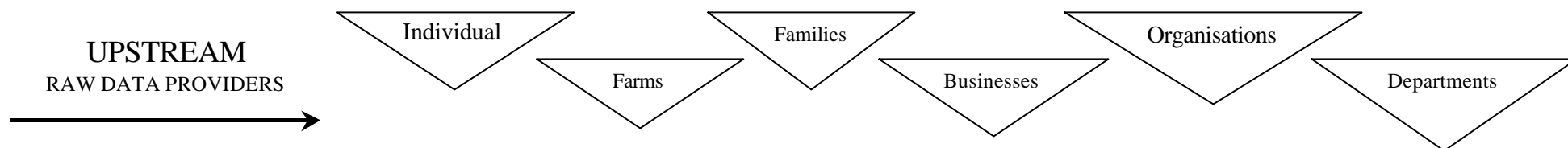
31. Notwithstanding these findings, ONS is currently re-examining its whole approach to the management of its data holdings and is looking afresh at the Data Warehousing option. Proposals currently being evaluated within the Office focus on the more traditional concept of a Warehouse as a system which embraces the entire statistical life cycle by connecting the upstream source systems to the downstream output systems. However, rather than tackling the warehouse concept using the 'Big Bang' approach, ONS is considering the possibility of implementing a much more limited 'pilot' exercise involving one or two data businesses in order to gauge the feasibility, practicability, and implications of implementing a much more comprehensive approach. The prototype which ONS is considering constitutes, in effect, a Data 'Mini-mart' rather than a Data Warehouse.

32. The imperative is to get it right before incurring the possibly huge costs associated with a full Warehousing approach. The business and cost implications revealed by this proposed 'proof of principle' experiment will not become apparent for some time to come. However, the expectation is that a well-planned Data Warehousing solution can bring GSOs the same sort of benefits which have been enjoyed by practitioners in the world of business:

- a world-wide shop window for their products leading, perhaps, to an expansion of their customer base
- greater multi-disciplinary coherence through opportunities to link up national and international information resources
- increased productivity through the more efficient exploitation of resources
- improved customer service and customer satisfaction and better interface with customers through on-line feedback procedures
- better market intelligence through the systematic tracking of customers' tastes and requirements.

TABLE 1

TRANSFORMATION PROCESS – RAW DATA TO INFORMATION



| PROCESS | LEVEL | ACTIVITY | INTERFACES |
|---|-------------|--|--|
| DATA TAKE-ON Interviews, Postal Questionnaires, Machine Extracts, etc. | Unit Data | Raw Data Capture and Storage | Provider Systems |
| PRIMARY PROCESSING | | e.g. Validity and Consistency Checks; Data Cleaning and Amendment Imputation | e.g. Records of Providers; Other Control Records |
| DATA STANDARDISATION AND DATA PACKAGING | Micro Data | e.g. Typological Referencing; Geospatial Referencing | e.g. Classification/Coding Systems (e.g. ISIC, ISCD, etc.); Geospatial Referencing Systems (e.g. Postcode Transformations, Spatial Co-ordinates, etc.) |
| SECONDARY PROCESSING | | e.g. Imputation; Removal of Discontinuities; Seasonal Adjustment; Weighting | e.g. XII |
| DATA AGGREGATION, EXPANSION, AMALGAMATION AND ANALYSIS | Macro Data | e.g. Time-series Tabulation; Cross-sectional Analyses; Information System Compilation | e.g. System Standards; National Accounts System; Balance of Payments System |
| DATA / METADATA DISSEMINATION | Information | Delivery through various hardcopy or electronic media | e.g. Style Guides; Format Guides; Language Conventions |



TABLE 2**THE DATA MANAGEMENT AND KNOWLEDGE MANAGEMENT PROCESS**

| ACTIVITY | INGREDIENTS | REQUIRED INTERFACES |
|--|---|--|
| Pre-Consultation and 'Peer Assistance' | Establishing, from the outset, the views and needs of peers, providers, and potential users in order to ascertain whether there is a genuine business justification for any proposed data collection, and no possibility of duplication of effort, and determine the best way forward. | Discussion Databases Data Archives Complementary systems Reporting Databases |
| Planning and Design | Activities associated with the design of the collection system or the content of the questionnaire in the light of earlier, complementary or other collection systems | Reference Databases containing, e.g. Sampling methodologies Question banks Etc |
| Harmonisation | Complying with national and international protocols, agreements, understandings, classifications and codes of practice relating to harmonisation, standardisation and integration in order to ensure full inter-operability of datasets | Reference Databases |
| Quality Assurance | Ensuring that all the various stages in the data life cycle meet appropriate methodological criteria, quality assurance standards and change control requirements in order to guarantee the statistical integrity of the dataset. Validating quality by reference to other collections | Reference Database containing e.g. Project Management tools Quality Targets Methodological Guides |
| Statutory Compliance | Fulfilling all relevant statutory or legislative obligations in order to ensure the security, confidentiality, legality, acceptability, and useability of the dataset, and safeguard any intellectual property rights attached thereto | Reference Database containing e.g. copies of relevant Statutes, Rules, Regulations, etc |
| Documentation | Full documentation of all the various steps in the data life-cycle, and retention and storage of any contextual and allied material, in order to preserve knowledge and expertise about the data system and its processes and outputs, maintain their functionality, and facilitate any subsequent data audits | Storage Database with documentation templates |
| Metadata compilation | Provision of easily accessible and comprehensive guidance material and interpretative text in order to foster awareness and understanding of the dataset and limit the possibility of misinterpretation | Ditto |
| Risk/Destruction Control | Action to ensure that the data are not inadvertently put at risk or corrupted at any stage in the life-cycle and only destroyed where it can be shown to providers, users and archiving bodies that the intrinsic value of the data is insufficient to justify the costs of preservation and retention | Back-up System |
| Archiving and Preservation | Where archiving is merited, using whichever managerial, organisational and methodological arrangements are most likely to guarantee the long-term preservation and re-useability of a dataset, taking into account recommended standards relating to the preparation of data in readiness for its preservation. | Archiving System |
| Dissemination | Dissemination of the data and accompanying metadata/documentation in whichever formats, and via whatever access routes will ensure that they reach the widest possible number of authorised users | Output Systems (using a variety of formats) |
| After-Action Review | Assessing what was supposed to happen, what did happen, why there was a difference, lessons learnt. | Reporting DataBase |

CHART A

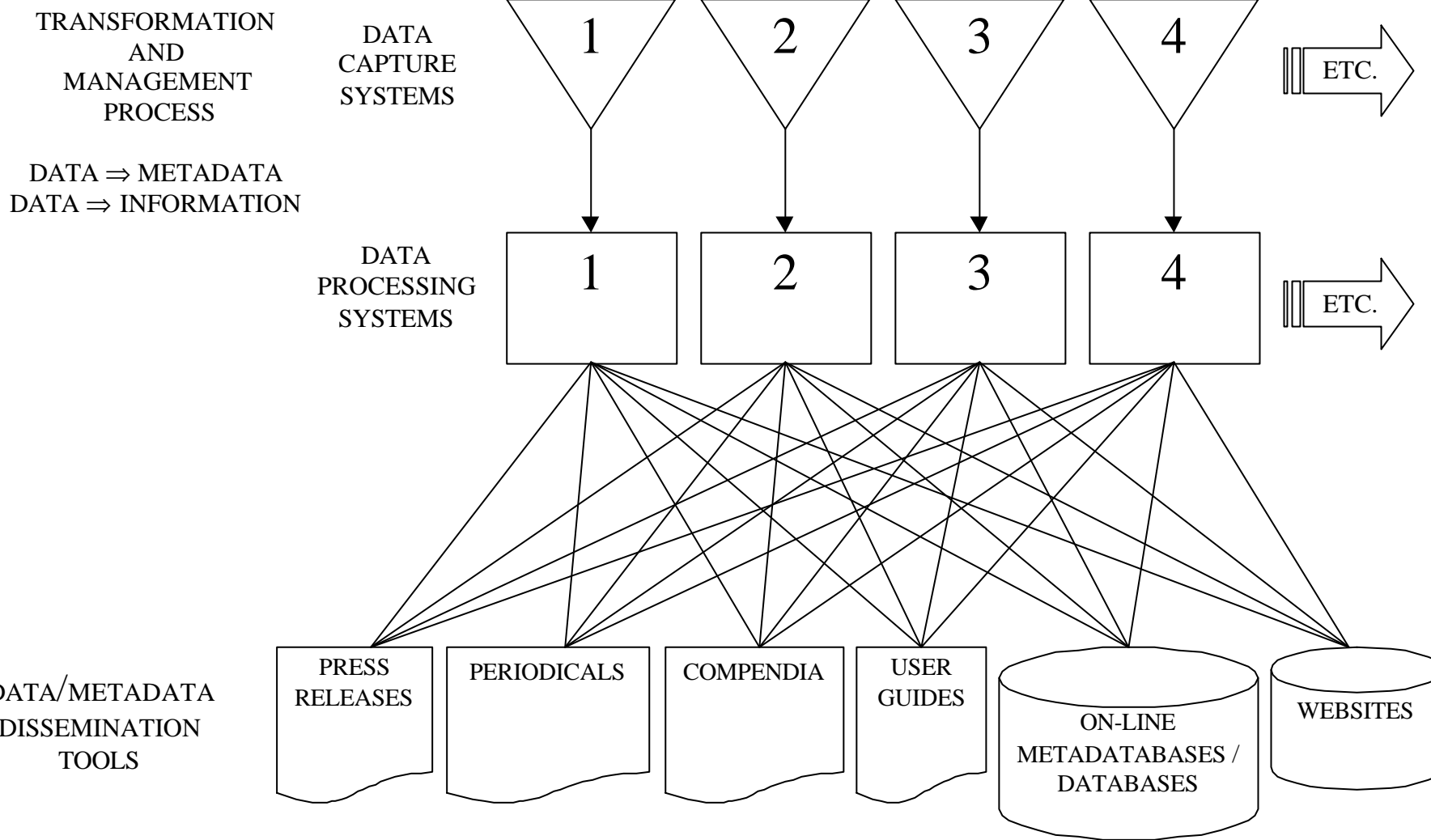


CHART B

TRANSFORMATION
AND
MANAGEMENT
PROCESS

DATA ⇒ METADATA
DATA ⇒ INFORMATION

DATA/METADATA
DISSEMINATION
TOOLS

