THE DESIGN AND DEVELOPMENT OF A SOCIAL SCIENCE DATA WAREHOUSE: A CASE STUDY OF THE HUMAN RESOURCES DEVELOPMENT DATA WAREHOUSE PROJECT OF THE HUMAN SCIENCES RESEARCH COUNCIL, SOUTH AFRICA

Andrew Paterson

Human Resources Development Research Programme, Human Sciences Research Council, Private Bag X41, Pretoria, 0001, South Africa
Email: anmpaterson@hsrc.ac.za

ABSTRACT

This article focuses on the development of a data warehouse to facilitate government decision-making on national human resources development and to provide public access to information. A set of key challenges was confronted in the development of the data warehouse including: the conceptualisation, design, implementation and management of the data warehouse system. The underlying questions that informed the process were, first: "In what ways will a data warehouse for a social science based research project be different from other database structures?" And second: "What are the particular management problems associated with large-scale long term social science based database projects?"

Keywords: Data Warehouse, Social Science, User Needs, Management, Integration

1 INTRODUCTION

This article focuses on the development of a database structure - called the "HRD Data Warehouse" - that is currently under construction by the Human Sciences Research Council. The aim of the Data Warehouse is to facilitate the provision of data and information about human resource development (HRD) in South Africa to stakeholders that include government, researchers and research agencies, and other interested parties in the public domain.

What is presented here focuses on a core set of issues that have been confronted in the process of developing the Data Warehouse. This account refers to conceptual, development and management processes and not to specific programming and software design aspects though these are obviously part of the challenge. Also, the focus of this paper will be on data warehouse management rather than on data management.

The development of the HRD Data Warehouse presents interesting challenges that inform the main discussion points in this paper. Fundamental questions addressed in the development of the Data Warehouse are:

- In what ways will the database[1] needs of a large-scale long-term (longitudinal) social science based research project be different from other databases?
- What are the particular management problems associated with a large-scale long term (longitudinal) social science based database project?

This paper will be structured in three sections. In the first section, a brief literature review is presented of the increased activity around data warehouse development in the social sciences. Then, the origins and aims of the HRD Data Warehouse project is presented. This is followed by an account of the increased prominence of human resources development in South African government planning that gave rise to the Data Warehouse project. Based on this background, an overview of the Data Warehouse project itself is provided.

The article then turns to answering the substantive questions raised in the introduction.

The second section is devoted to an exploration of the differences and similarities between the HRD Data Warehouse, enterprise data warehouses and data archives. This is undertaken in order to shed some light on the extent to which the HRD Data Warehouse represents a unique attempt to meet challenges in the dissemination of social science data.

In the third section, important management problems associated with the HRD Data Warehouse are discussed. This is followed by a brief concluding section.

2 DATABASE AND DATA WAREHOUSE ACTIVITIES IN THE SOCIAL SCIENCES

Information and communication technologies provide the medium for the generation, storage and analysis of an ever growing universe of digital information. This is clear from research in the physical and life sciences sectors which are information intensive, generating huge volumes of data and presenting significant requirements for data storage, management, integration, processing and analysis (Milburn, 2001; Martin, 2001; Mack & Hehenberger, 2002, PTC, 2002; Bizzozero, 2002).

The application of database technologies to facilitate social and human science research and research dissemination is also increasing. The following examples are given of the range of different database applications for structured and unstructured data in the social science field:

- archiving of qualitative data (Corti & Thompson, 1997) and primary historical data (Walker, 2002)
- cross-national database integration (Tanenbaum & Mochmann, 1994; Sinott, 1994; Kondro, 1999)
- web-based publication and dissemination of data (Bainbridge, 1999)
- development of natural language database search aids for the humanities (Knapp, 1998) and lexical approaches to textual data analysis (Bolden & Moscarola, 2000)
- development of social science Geographical Information Systems (Bainbridge, 1999)
- data mining and social science data exploration and description (Hand, 2000; Korth & Silberschatz, 1997)
- creation of an open source digital library system for quantitative data that provides the infrastructure for the dissemination of distributed collections of quantitative data (Altman, Andreev, Diggory, King, Sone, & Verba *et al*, 2001)
- establishment of a federation of distributed interoperable anthropology databases (Digital Archive Network for Anthropology) (Clark, Slator, Perrizo, Landrum III, Frovarp, & Bergstrom *et al*, 2002)
- Large international consortia such as the Council of European Social Science Data Archives (CESSDA) and the Inter-University Consortium for Political and Social Research (ICPSR) have emerged to support the storage and dissemination of social science data.

Database and data warehouse structures provide the platform for increased collaboration between researchers and for improved visibility and access to social science information in the public domain. Government is a major consumer of social science research which can inform policy making, monitoring and evaluation of delivery on the ground.

In developing countries information sharing and dissemination is viewed as a potentially valuable resource that can feed into social development projects and support sustainable development activities in rural contexts and among marginalized communities (Funk, 1999). Furthermore, the literature shows an increased stress on the importance of improving science communication in order to raise public understanding of science and technology issues (Logan, 2001; Weigold, 2001; Treise & Weigold, 2002).

The development of the HRD Data Warehouse in South Africa is reflective of the trend towards the application of database structures in the social sciences in support of government development goals and in order to improve public access to information. The following sections of this article will briefly describe the aims and key characteristics of the HRD Data Warehouse.

3 THE AIMS OF THE HRD DATA WAREHOUSE

In October 2000, the South African Cabinet appointed the HSRC as the official agency responsible for undertaking and providing government with research support services in the field of HRD. Planning and co-ordination of HRD has become a top government priority in the era of globalisation with its associated 'knowledge economy'. Effective HRD policies are a key requirement of successful co-ordinated market economies that optimally balance conditions prevailing in the education and training system, the labour market and the economy. HRD is therefore a cross-sectoral policy issue. It impacts on a multitude of policy domains such as education and training, labour market, macro-economic, industrial and foreign trade (DACST, 1996; Department of Labour, 1997; Department of Trade and Industry, 1998). In the absence of proper cross-sectoral co-ordination, policies would remain isolated and less effective.

In this context the implementation of an appropriate ‘management information system’ becomes a prerequisite. A key step in the development of HRD policies is the acquisition of detailed quantitative and qualitative descriptions of conditions prevailing in the education and training system, the labour market and the economy. Such information will be essential to capture and monitor the interaction of policy effects in the domains relevant to HRD. The task of quantitatively describing the phenomena associated with HRD in any national economy is significant. This is because of

the complexity and scale of the data potentially relevant to understanding and tracking change in HRD, which is inherently a research driven process with a large data component. An unrestricted approach to data acquisition for the HRD data warehouse would have been unviable, because of the likelihood that data of marginal relevance to core HRD themes would be collected leading to wasteful expenditure on seldom used databases. In order to avoid this danger, a framework was developed in order to guide the selection of data to be included in the warehouse.

An important characteristic of the HRD Data Warehouse is that a theoretical model informs what data is deemed relevant for inclusion. The collection of data is targeted and informed by a coherent theoretical model of the complex dynamics at work in human resources and labour markets in the South African economy. This enabled the identification of dimensions most critical to understanding HRD in South Africa. Twenty-five key dimensions were identified, which cover both demand and supply sides of the labour market. For each dimension, appropriate measures - sometimes indicators - were selected for the purpose of recording and tracking the social and economic phenomena relevant to HRD, and for monitoring system change towards stated policy goals.

Thus the HRD Data Warehouse could be said to be unique because its structure is based on a theoretical model of the domain of human resource development. Other social science research data warehouse and archive facilities do not necessarily have such an overarching framework, and as a consequence will collect a wider range of databases which may not necessarily bear on the same social issue or problem.

4 MAIN FUNCTIONS OF THE DATA WAREHOUSE

4.1 Research reports data tables and databases

As noted above, the theoretical model for the Data Warehouse recognises twenty five key dimensions. Accordingly, academic and research experts were commissioned to write on each of the twenty-five dimensions. Their work forms the basis of the Data Warehouse. Each researcher will source primary data, conduct analyses and provide their data and analyses for publication on the HRD Data Warehouse website. The Data Warehouse will house the chapters commissioned from the expert researchers, as well as all the accompanying data tables and databases that they submit. The data submissions will correspond to the stages of data organisation, of aggregation, and of refinement that support the final analysis presented in each research report. Thus it will be possible to replicate analysis at a later stage.

4.2 Acquisition of data

The Data Warehouse is designed as a long-term project because changes in education and labour market systems are slow. Databases will accumulated through regular phases of data collection for longitudinal analysis by commissioned researchers, as well as through the ongoing sourcing of other relevant databases on an ad-hoc basis. Consequently, the data warehouse is designed to support the assembly, storage, manipulation and analysis of a potentially large number of sizeable databases.

The process of accumulating databases in the warehouse will be selective. Such databases must be relevant to the definition of the HRD research field and must meet quality (validity, reliability and compatibility) criteria.

The aim is to acquire sub-sets (specific fields or variables) of government databases, rather than to go into full-scale duplication of government data. In this way, unnecessary and uneconomical accumulation of data will be guarded against. Also, the responsibility for maintenance and updating of the parent databases will remain with the source/originating government departments such as: the Department of Labour, the Department of Education, the Department of Trade and Industry, the Department of Agriculture, and the Department of Science and Technology.

The main types of data that will be held in the Data Warehouse are:

- Key databases that have been obtained from either government sources and which are located in the HRD data archive
- Additional databases will be created in the course of ongoing research that the HSRC will conduct as part of the HRD research programme
- Additional databases from independent external sources, including tertiary institutions, independent research entities and non-government organisations. Obtaining these data sets is important because such existing data has the potential to add value to the research activities associated with the HRD project – at minimal additional cost.

4.3 Data tracking system

The data tracking system makes it possible to trace the evolution of databases – or of the data tables drawn off them - from their raw form through to final presentation form, so that the process of data manipulation can be audited and checked.

4.4 Meta-data catalogue

A meta-data catalogue will assist effective identification and retrieval of information. It will also enable users to determine which data may suit their needs, and will refer researchers to related data for comparative purposes (El Sherbini, 2001). The metadata catalogue contains information about each database source including: information on the data itself (eg: size of dataset, format, codebook, GIS integration, dimensions/fields, unit of collection/aggregation, etc.), confidentiality, conditions for release, frequency of collection etc. The Dublin Core metadata standard has been adopted as the basis for the development of the meta data catalogue.

4.5 Data access

Data and information will be stored as follows:

a) Research reports (in PDF or HTML) which are derived from the data resources in (b), (c) and (d) below:
b) Data represented visually in the form of graphs, charts, maps, summary tables etc
c) Data tables (in MSExcel)
d) Databases

The client is able to download web-enabled data and information from (a) to (c). The database files in (d) can be supplied to the client upon request and subject to permissions from the source.

The data in (b) and (c) are searchable, but the data is pre-formatted so there will not yet be the opportunity to query the data dynamically. Access to the data tables is limited to drilling down and rolling up through tables layered by level of aggregation. The website also provides access to supporting documentation, metadata, and other relevant research documents to assist clients in their own analysis as needed. With time, functionality will be developed to include more complex data exploration and analysis techniques.

5 THE DATA WAREHOUSE IN COMPARISON WITH THE ENTERPRISE DATA WAREHOUSE AND THE DATA ARCHIVE

The system requirements and functionality described above prompted a search to establish the existence of models of similar projects. However, it was discovered that the HRD Data Warehouse is a hybrid type of system that shares similar features with at least two recognizable database systems. They are the research data archive and the enterprise data warehouse.

A comparison between these systems was conducted in order to generate clarity on what the main features of the envisaged Data Warehouse should be. This exercise revealed that each database type has certain unique characteristics. A summary of the comparison is presented in Table 1 below. The dimensions according to which the database types were compared are as follows.

- First, the *underlying institutional context*, since the use value of a data warehouse to an organisation should serve as the primary consideration shaping the design.
- Second, the *primary client base*, because this group will need to use the data
- Third, the *processes driving the collection of data*, because this would determine the nature and quality of the data.
- Fourth, the *extent to which data integration can be implemented*, because in large-scale database configurations, higher levels of data integration enables automation of querying, data mining and reporting. Low levels of data integration increase the requirement for human analysts to perform analysis
- Fifth, the *types of analysis enabled* by each database configuration was considered since the type of analysis must have the potential to inform individual or institutional action

There are obvious similarities between the three. They all are potentially large physical databases holding potentially vast amounts of read-only data from a variety of sources, where the data is non-volatile, is not connected to operational systems and is to some degree organized on a subject oriented basis (Inmon, 1992; Robb & Coronel, 2000). But this is where the clear similarities end.

The major distinguishing characteristic of the archive type is that submission of data is voluntarist within a loosely bound social science community. As a result of this dependency, there is minimal control over what data is collected or

the underlying methodologies, since this is the prerogative of individual social scientists. As a consequence, databases submitted will not necessarily bear any relation to each other, thus ruling out integration.

The HRD Data Warehouse does share some similarity with the enterprise data warehouse, in that both are focused on obtaining data for a system to support decision making. But the characteristics of the systems are very different. The enterprise data warehouse is a closed system, where designers can maximize their control over the nature and quality of data gathered which is focused very specifically on enterprise goals. The potential for data integration and automation of queries is greatest in a controlled and focused design environment.

In contrast the South African national human resources development (HRD) domain is an open and highly complex system with social, economic, political and psychological dimensions. The scope of the HRD domain, the complex interrelationships between forces and actors, and the methodological challenges in quantifying social change present considerable challenges which demand human interpretive intervention.

It is clear that client needs are a significant factor. In the case of the enterprise data warehouse client definitions are least problematic since these are at least defined within the boundaries of the organisation. In the case of the data archive this presents some difficulties, since the definition of who the clients are could only be understood with reference to disciplinary boundaries but even these are increasingly permeable. The difficulty for the HRD Data Warehouse is that it seeks to meet the needs of a diverse client base from government officials to the general public.

Table 1. Comparison between an enterprise data warehouse, the HRD Data Warehouse and a data archive

	ENTERPRISE DATA WAREHOUSE	HRD DATA WAREHOUSE	DATA ARCHIVE
CONTEXT			
Institutional Base	Single enterprise	South Africa National Human Resources Development field. Multiple government institutions	Academic association. Voluntary membership
Aim	Competitive advantage of enterprise in market	Coordination of HRD for national competitive advantage and improved social equity	Centralisation of data and information resources for the purpose of sharing
Guiding Framework	Corporate strategy	Legislation, regulation and policy	Informal commitment to community of practise
CLIENTS			
Client base	Enterprise management	Heterogeneous -Government managers, researchers and general public	Open, but limited by interest in research, disciplinary base and methodological expertise
DATA			
Schedule of data gathering	Frequent and regular (dependent on business model). Based on management needs of enterprise and need for fast response times	Replication on a biennial or triennial basis. Social change is slow	Mostly once-off data sets
Data standards	High levels of control	Specified to researchers but dependent on sources of data	Limited prior control. Quality assurance upon submission only
Acquisition of data	By specification of data required for management decision making	By commissioning of researchers through written briefs	By voluntary submission
DATA INTEGRATION			
Data source(s)	Limited and internal to the enterprise	Multiple	Multiple
Links between databases	High possible levels of relatedness	Limited	No necessary links
Organisation of data	In a relational database	By theme as specified in the HRD conceptual model	Archived by field of study and or methodology and discipline
Level of data integration	High	Low levels of integration limits analysis based on querying the database.	None expected
ANALYSIS			
Analytic method	Reports from queries via software technology	Mostly discursive via human analysis	No analysis between databases

6 WORKFLOW MANAGEMENT FOR THE HRD DATA WAREHOUSE

In this section, the workflow of the HRD Data Warehouse (Figure 1 below) is briefly discussed in order to identify how this differs from the enterprise data warehouse and archive types.

First there is considerable emphasis on controlling the flows of data into and out of the HRD Data Warehouse system. Data can be sourced through two channels, namely through research undertaken by commissioned researchers or from the data holdings of various government agencies (eg: Department of Education (DoE), Department of Labour (DoL), Department of Arts Culture Science and Technology (DACST) and Statistics South Africa (StatsSA)). Data dissemination out of the HRD Data Warehouse must meet different client needs from government managers to researchers to the general public.

In contrast, the enterprise data warehouse sources all data from internal systems, while the data archive receives rather than sources data. It is therefore important for the data warehouse system to adequately govern the system of data flows especially in the case of government data that may have restrictions on ownership and publication. Furthermore, in all instances where there is sourcing of data from external sources, attention must be given to ensuring acceptable levels of data quality.

It is clear that the Data Warehouse requires additional processes of data management at the interface with suppliers and sources of data. This necessitated the establishment of a Data Committee and the drafting of agreements or Memoranda of Understanding (MOU) with the participating government departments. In contrast, the enterprise data warehouse, which sources its own data internally, does not have the additional management requirement.

A critical issue upon which the success of the HRD Biennial Directory and Data Warehouse rests is the quality control of data used for interpretation in the research outputs. The Data Warehouse is intended as an authoritative source of information and data on HRD in the years to come and the validity of its data should be highly rated. The quality and validity of interpretation of data must be carefully controlled because of the need to outsource work to researchers external to the HSRC, and consequently, the high number of research participants. Hence, a quality control mechanism will have to be put in place to ensure compliance with standards applied in the Data Warehouse.

With regard to the communication of data, the HRD Data Warehouse disseminates data to external communities via a web interface, whereas an enterprise data warehouse may web-enable its data, but this will normally accessible only internally to the organisation.

The HRD Data Warehouse is part of a dual strategy to disseminate research findings. The development of the data warehouse is taking place parallel to the publication of a “Biennial Directory” which will contain research reports on each of the twenty five dimensions of the HRD domain identified in the theoretical model. In this way, an interactive relationship will be established between the hard copy publication of the Biennial Directory and the digital resources in the Data Warehouse. For example, it is envisaged that the reader of the Biennial Directory will be able to find additional data on an issue of interest that could not be included within the constraints of the published Biennial Directory. This represents an experiment in hybrid, or dual channel – print and digital – dissemination of information.

The discussion will now shift to an analysis of key challenges encountered in the development of the HRD Data Warehouse.

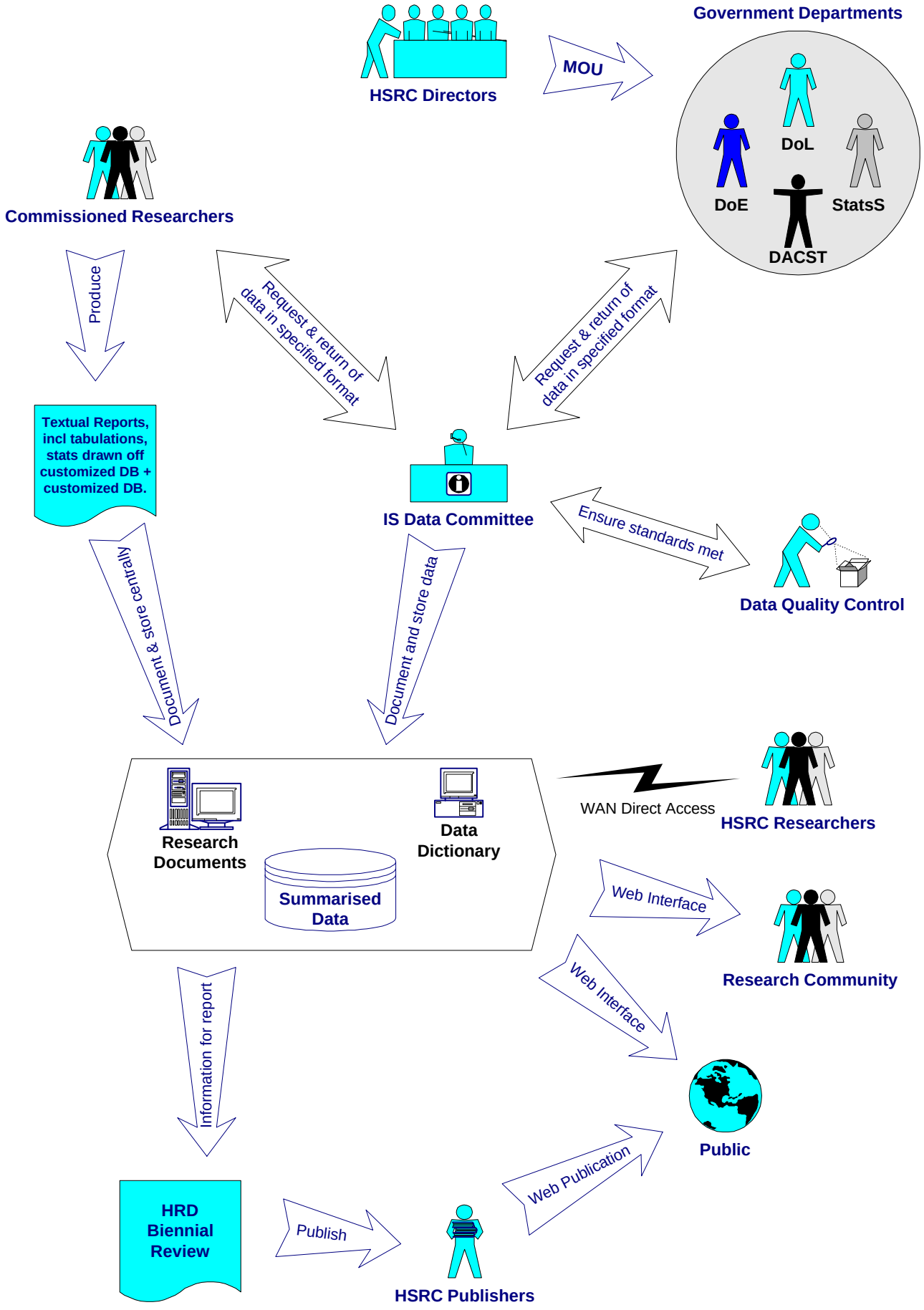

Figure 1. Workflow of the HRD Data Warehouse

7 INTEGRATION

It is important to understand what 'integrated database system' means in the context of the Data Warehouse development. The term 'integration' seems to be used with different meanings and with different purposes. It connotes the following meanings:

- One single technical platform for database, interface and development tools
- A complete business 'solution' offered by a vendor
- A homogenous user interface
- That systems have to be compatible in terms of content, terminology and their ability to exchange data.

It is with the latter meaning that this paper is concerned. This is because the idea of obtaining data compatibility is very attractive for the social scientist. In the ideal, this would mean that a researcher could compare data across related phenomena. To take an example from the HRD project, a researcher may wish to compare data on unemployed school leavers with data on the informal economy to see if levels of informal economic activity increase with levels of unemployment. In order for this to be possible, the two data sets would have to be compatible:

- The spatial unit of analysis would have to be the same ie: it would be of no use to have one set of data only available at the provincial level and the other at the national level.
- The timing of data collection would have to coincide as closely as possible or the time elapsed between the time of data collection will produce validity problems
- The definitions of 'school leaver', 'unemployment' and 'informal economic activity' would have to be exclusive. Otherwise the data will reflect perspectives of the same phenomenon.
- The data itself must be technically compatible.

These conditions cannot easily be achieved. This is because different information sources will differ in their internal data structure (file systems, relational databases etc), protocols for the treatment of information (how it is recorded) as well as in organizational (classification) and technical (software and hardware) respects.

For the HRD project, substantial difficulties with integration of data from various sources are noted. Those difficulties increase with the number of data providers over which the Data Warehouse has no control. On the basis of there being multiple independent sources of database creation and supply, it stands to reason that there is no standardized data management strategy between the data suppliers.

Therefore in the process of planning social science database integration, such conditions leading to the type of structure and content of databases need to be understood, so that the basic requirements for integration and the means of harmonization can be identified (Di Cesare & Lazzari, 2000). Even approaching the ideal of integration is difficult, given the diversity of producers of data, the fragmentation of information sources and the consequent absence of - or unstandardised - taxonomic and classificatory control. For example, entities (eg: schools, districts, provinces) will have multiple attributes, coding schemes, and values across databases.

Martin & Leben (1989) observe that as a consequence of a lack of required standards, the establishment of linkage and integration may require "excessive time and human resources". This observation is particularly true for the HRD Data Warehouse project.

So, the difficulty of integration on a technical level has to be overcome at another level. The integrating source in the Data Warehouse lies in the actual theoretical framework developed, which relates the different domains of HRD into a unified explanatory framework. This is naturally not complete or systemically driven on account of the complexity of the HRD field and also because HRD is not a closed system. Where programmatic or technical solutions to data integration reach their limits, the integrative value in the project derives from the analytic input of the researchers and the way they relate the findings to other domains of HRD.

8 OWNERSHIP OF AND CONTROL OVER DATA

The flows of raw data and of data products in and out of the Data Warehouse present particular challenges with regard to the protection of ownership and confidentiality of data. One important difference between the HRD Data Warehouse and an enterprise data warehouse is that the latter is a relatively closed system with a defined set of users. By contrast, the HRD Data Warehouse has multiple external suppliers and users of data, which creates the need to devise multiple procedures to deal with different requirements. This is particularly true of a social science database infrastructure, which contains data that is comparatively recent – i.e. is not historical – and as a consequence can contain information about individuals or institutions that may be considered sensitive, or which may yield research results that are controversial.

First, research on HRD requires access to data, where confidentiality is of concern. For example, a participating government department may require that data below a certain level of disagregation may not be made publicly available so that the identity of institutions and individuals is protected. Such restrictions should be honoured, whatever limitations this may impose on the dissemination of data from the Data Warehouse.

Second, organisations from which the Data Warehouse obtains data may impose conditions on the use and dissemination of the data. For example, an organisation may stipulate that it will provide data to the Data Warehouse on condition that only the HSRC can conduct analysis and publish results, and that the database should not be made publicly available. Such an arrangement could fall foul of the Constitutional guarantees of access to information. Another type of condition is where data may be provided to the HRD Warehouse only on condition that the policy of the Warehouse is to supply data to third parties free of charge.

Third, the HRD project as a whole is intended to have a long-term trajectory. The aim is to track change in national human resources development over time on a longitudinal basis, which means that methodologies must be applied consistently over time for the production of key indicators. For time based socio-economic analysis to work, it is important to ensure that procedures for data collection and analysis are standardised and then adhered to for each iteration of the research. It is therefore necessary to record the data manipulation procedures, assumptions and calculation methods initially used. There are good reasons for this. The possibility exists that at some point in the future, it may be necessary to return to old datasets in order to recalculate indices or indicators for one of the HRD dimensions because of changes in the phenomenon under scrutiny. Alternatively it should be possible to replicate procedures with the relevant dataset in case there are any queries concerning the reported results.

However, researchers may view the techniques they apply in cleaning and improving datasets, and in analyzing data as their own value added intellectual property. Their concern is that they would lose control over this embedded value if it were released into the public domain. This raises the question as to how to deal with intellectual property rights that reside in databases. In certain instances, the HRD Data Warehouse may need to undertake that neither the raw data, nor the methodology that the researchers use, will be made available to any other person(s) or organisations. The data and methods are kept strictly confidential other than for application in ensuing versions of the directory.

The aspects noted above necessarily restrict the Data Warehouse from making its methods public, which is in contradiction with the idea of a transparent approach to strategic data analysis and dissemination. Such a tension, it seems, is inevitable when handling social science data.

9 CHOICE OF PROTOCOL FOR THE DEVELOPMENT OF SYSTEMS SUCH AS THE DATA WAREHOUSE

Frameworks for building information systems usually specify a methodology. This consists of a body of proven knowledge and techniques for undertaking a repeated task and which can produce more or less reliable outcomes.

There are a variety of frameworks that can be used to guide or structure the process by which information systems are built. The main characteristics include the following:

- Some are more formal than others.
- Some methodologies have a stronger user-centred approach while others confer much more decision making power to the technical ‘experts’ such as systems developers/designers/architects.
- Some are more people oriented while others are more process oriented (Fowler, 2001).
- Some systems methodologies are more prescriptive while others provide rules for contingent factors in the site where the system is to be developed (Middleton, 2000).
- Some methods are more predictive and less adaptive to change (Fowler, 2001).

The different methodologies should not be seen as being in competition, but rather as complementary for a number of reasons. There are many different information systems in different organisational environments, with different purposes and at different stages of development or maturity as described by the Gibson-Nolan model of EDP growth (Modell, 2000).

The various methodologies can be understood along a continuum from:

(a) Classical sequential or “waterfall” methodologies that draw their inspiration from the disciplines of physical engineering (eg: construction, civil, mechanical etc). An example of a framework is the Systems Development Lifecycle (SDLC)

To the

(b) Agile (sometimes called 'lightweight') methodologies that have an adaptive and people-oriented approach (eg: these range from the better-known XP (Extreme Programming) to the Rational Unified Process (RUP) (Fowler, 2001).

There are obviously many variations of the above that meet the needs of particular users. For example:

- The UK Government has its own mandatory structured systems analysis and design method (SSADM) for all government IS (Hutchings, 2002)
- RAD or "Rapid application development" methodologies that were developed in an unstructured way. This produced the impetus for joint development of a public domain RAD method – the Dynamic Systems Development Methodology (DSDM) (DSDM, 2002)

In the context of the HRD Data Warehouse development, it was clear that a highly structured and hierarchical model for systems development was not applicable for a design process with a wide set of possible participants including: data providers, research teams, government clients, research clients and public users, as well as information professionals, web developers, programmers, systems administrators and content developers. The design of the Data Warehouse was therefore progressively developed through a recursive set of stages leading to the final structure.

10 CONCLUSION

In the attempt to understand the nature of the HRD Data Warehouse, this article explored the differences between the HRD Data Warehouse, a 'typical' model of a enterprise data warehouse and a 'typical' model of a data archive. In so doing, it exposed some of the key challenges associated with assembling social science data in a coherent fashion. The outcome of the discussion suggests that the HRD Data Warehouse is unlike a data archive for its attempt to focus on a particular social and economic domain. The domain in question is constituted in a set of open and interrelated social educational and economic sub-systems and activities. This means that what is considered important or relevant to understanding human resources development has to be defined through the formulation of a theoretical model, which must necessarily prioritise particular phenomena – and their associated data – for explanatory purposes. In contrast, designers of an enterprise data warehouse select data for capture on the basis of its being mission critical to sustaining competitive advantage. The bounded nature of enterprise information systems in combination with the high levels of control over data selection makes it possible to construct a database system with high levels of integration.

Similar levels of data integration may not be practically attainable within a social science environment. The cost of obtaining social scientific data on a systemic level is an activity usually undertaken by governments because of the costs associated with such large undertakings. The underlying rationale and methodology for collecting data, is determined by government in accordance with its own needs. This means that a social science data warehouse which seeks large scale data will not usually have the opportunity to inform the process of data prioritisation or collection and must utilise data as it is received. Available data may not be amenable to technical integration. Therefore the selection of which data to be used in analysis must be a theoretically informed interpretive task.

The challenges associated with the development of a focused social science data warehouse therefore lie in obtaining the best combination of automated querying and analysis of data obtained from external sources – such as government – with informed discursive analysis. In other words, analysis of data must be informed by theoretical framework which will be undertaken by researchers.

Finally, this article suggests that the evolution of a focused social science data warehouse must entail significant levels of interaction between clients, data suppliers, researchers and system designers. Such complex levels of interaction preclude the planning of the system within the parameters of a structured approach to information systems design and development.

The tendency of organisations to harvest and to store data in ever-greater quantities continues. A corresponding increase of the same magnitude in our capacity to identify, make sense of and disseminate appropriate data seems to be in the offing. The ways in which data and database structures are made visible and accessible to clients are becoming increasingly important – not least for the HRD Data Warehouse. As database and data warehouse architectures and configurations become more successfully aligned with the purposes of the business in which they are located, we will see increased hybridisation taking place. Database and data warehouse systems are beginning to look very different from each other as they 'fit' the needs of the organisations and client information needs which they must serve.

11 ACKNOWLEDGEMENTS

The development of the HRD Data Warehouse system has taken place within a team. The immediate team is: Robin Naude, Arjen Van Zwieten and Andrew Paterson with Andre Kraak, Eberhard Kobler, Helen Perry, Gerald O'Sullivan, Faye Reagan and Corporate Services IT. In addition, the process was informed by inputs from the South African Departments of Labour, Education, and Science and Technology. Thanks to Robin Naude who constructed the workflow diagram, to Arjen Van Zwieten and Heidi Paterson for their careful reading and comments on the drafts of this article, and to Zubeida Tayob for administrative assistance. Finally, thanks are also due to two anonymous reviewers for their constructive and valuable criticisms of earlier versions of the article. The development of the HRD Data Warehouse is funded by the South African Department of Science and Technology

12 REFERENCES

Altman, M., Andreev, L., Diggory, M., King, G., Sone, A., Verba, S., Kiskis, D.L., & Krot, M. (2001) A digital library for the dissemination and replication of quantitative social science research. *Social Science Computer Review 19*(4), 458-470.

Bainbridge, W.S. (1999) International network for integrated social science. *Social Science Computer Review 17*(4), 405-420.

Bizzozero, S. (2002). Life sciences are alive and kicking. *Guardian Unlimited*
Retrieved November 14, 2002 from the World Wide Web:
www: http://education.guardian.co.uk/mbas/stpry/0,10671,743847,00.html

Bolden, R., & Moscarola, J. (2000) Bridging the quantitative – qualitative divide: The lexical approach to textual data analysis. *Social Science Computer Review 18*(4), 450-460.

Clark, J.T., Slator, B.M., Perrizo, W., Landrum III, J.E., Frovarp, R., Bergstrom, A., Ramaswamy, S., & Jockheck, W. (2002) Digital archive network for Anthropology. *Journal of Digital Information 2*(4), 1-15.

Corti, L., & Thompson, P. (1997) Latest news from the ESRC Qualitative Data Archival Resource Centre. *Social History* 22(1), 83-87.

Department of Arts, Culture, Science and Technology (DACST) (1996) *White Paper on Science and Technology: Preparing for the Twenty-First Century*. Pretoria: Government Printer.

Department of Labour (1997) *Green Paper on a Skills Development Strategy for Economic and Employment Growth in South Africa.* Pretoria: Government Printer.

Department of Trade and Industry (1998) *Industrial Policy and Programmes in South Africa: Discussion Document.* Pretoria: Government Printer.

Di Cesare,R., & Lazzari,G. (2000) Towards integration of information sources on grey literature: A case study. *International Journal on Grey Literature 1*(4), 167-173.

DSDM (Dynamic Systems Development Method) (2001) The history of DSDM.
Retrieved November 28, 2002 from the DSDM Consortium website:
http://www.dsdm.org/en/about/history.asp

El-Sherbini,M. (2001) Metadata and the future of cataloguing. *Library Review 50*(1), 16-27.

Fowler, M. (2001) The new methodology.
Retrieved November 28, 2002 from the Martin Fowler website:
http://www.martinfowler.com/articles/newMethodology.html

Funk, K. (1999) Information networking as an instrument of sustainable development. *Social Science Computer Review 17*(1), 107-114.

Hand, D.J. (2000) Data mining. *Social Sciences:* New challenges for statisticians *Social Science Computer Review 18*(4), 442-449.

Hutchings, T.D. (2002) Introduction to methodologies and SSADM.
Retrieved November 28, 2002 from the University of Glamorgan, School of Computing website:
http://www.comp.glam.ac.uk/pages/staff/tdhutchings/chapter4.html

Inmon, W.H. (1992) Data Warehouse – a perspective of data over time. *Data Base Management* (Feb), 370-390.

Knapp, S.D., Cohen, L.B., & Juedes, D.R. (1998) A natural language thesaurus for the humanities: the need for a database search aid. *Library Quarterly 1* (October), 1-15.

Kondro, W. (1999) Making social science data more useful. *Science 286*(5441), 880.

Korth, H.F., and Silberschatz, A. (1997) Database research faces the information explosion. *Communications of the ACM 40*(2),139-143.

Logan, R.A. (2001) Science mass communication. *Science Communication 23*(2), 135-163.

Mack, R. & Hehenberger, M. (2002) Text-based knowledge discovery: Search and mining of life-sciences documents. *Drug Discovery Today 7*(11), S89 – S98.

Martin, A.C.R. (2001) Can we integrate bioinformatics data on the Internet? *Trends in Biotechnology 19*(9), 327-328.

Martin, J. & Leben, J. (1989) *Strategic information planning methodologies.* Englewood Cliffs, NJ: Prentice Hall.

Middleton, P. (2000) Barriers to the efficient and effective use of information technology *International Journal of Public Sector Management 13*(1), 85-99.

Milburn, J. (2001) Beyond the genome: Turning data into knowledge. *Drug Discovery Today 6*(17), 881–883.

Modell, M.E. (2000) The various types of information systems analysis projects
Retrieved November 28, 2002 from the Dai-Sho Inc. website: http://www.dai-sho.com/pgsa2/pgsa02.html

PTC (Pittsburgh Technology Council) (2002) Capitalising on the convergence of the life sciences and IT *Technology Entrepreneurship Quality 8*(5), May. Retrieved November 14, 2002 from the World Wide Web:
http://www.pghtech.org/news/teq/teqstory.cfm?id=786

Rob, P., & Coronel, C. (2000) *Database systems: Design, implementation and management.* Cambridge, MA: Thompson Learning.

Sinott, R. (1994) Theories of integration and the integration of the European database. *International Social Science Journal 46*(4), 533-541.

Tanenbaum, E., & Mochmann, E. (1994) Integrating the European database: Infrastructure services and the need for integration. *International Social Science Journal 46*(4), 499-512.

Treise, D., & Weigold, M.F. (2002) Advancing science communication. *Science Communication 23*(3), 310-322.

Walker, K. (2002) Integrating a free digital resource: The status of Making of America in academic library collections. *RLG DigiNews 6*(1), 6-12.
Retrieved 25 February 2002 from the World Wide Web: http://www.rlg.org/preserv/diginews/diginews6-1.html

Weigold, M.F. (2001) Communicating Science. *Science Communication 23*(2), 164-193.

ENDNOTE

[1] The term ‘database’ is used to refer to a collection of information that uses a variety of media such as IT databases, other electronic media (for example: MSWord documents, MSExcel spreadsheets, JPEG graphs, CD ROMs etc) and printed documents. A catalogue or index (data dictionary) with a search engine will ensure that such a collection of information becomes a ‘system’.