A Corporate Data Repository For CCLRC Using CERIF

EDDY GRABCZEWSKI, SHIRLEY CROMPTON, STUART ROBINSON, TRUDY HALL
CCLRC

Summary

This paper documents the history, requirements and work in progress to develop a Corporate Data Repository (CDR) in support of three sites owned by the *Council for the Central Laboratory of the Research Councils* (CCLRC). From the outset, the CERIF standard has been our starting point in developing a common Corporate Data Model (CDM). To support our business processes we have extended the CERIF data model in ways that are described in this paper. Our design philosophy is to develop a data model that is upwards compatible with CERIF. This paper demonstrates that the CERIF standard can be incorporated not only into CRIS applications but also into corporate information systems such as that at CCLRC.

1 Introduction

The *Council for the Central Laboratory of the Research Councils* (CCLRC) is an independent, public body of the *Office of Science and Technology*, which is itself a part of the *Department of Trade and Industry* in the United Kingdom. Formed in 1995, CCLRC owns and operates three sites: Rutherford Appleton Laboratory (RAL), Daresbury Laboratory (DL) and the Chilbolton Facility (CF). These sites support advanced research for the worldwide scientific community which include *ISIS* - the world's most powerful neutron and muon source, *CLF* – a state-of-the-art laser facility, *SRS* – the world's brightest UV and X-ray light source, *HPCx* – one of the worlds most powerful computers together with other equipment such as radio telescopes, a modern electron microscope and other engineering facilities to build space science equipment. The *Business and Information Technology Department* (BITD) supports these facilities and services.
To update this complex organisation, CCLRC decided to integrate its disparate information systems. This paper maps the work in progress to develop a Corporate Data Repository (CDR) to support the three sites. The CERIF standard (euroCRIS, 2003) has been fundamental to our approach in developing a common Corporate Data Model (CDM) for CCLRC.

2 History of the CDR

Between 2001 and 2002 the CCLRC Business Systems Strategy Group (BSSG) identified and agreed an overall strategy to "*Improve the design and operation of CCLRC's business processes, thereby increasing the effectiveness, job satisfaction and motivation of staff, the quality of the delivered services and more effectively supporting the attainment of CCLRC's mission*" (CSSB, 2003). To achieve this strategy, the following objectives and tactics were employed:

Objective: to automate all business processes.
Tactics:

- Identify existing business processes by examining existing paper and electronic forms.
- Identify the underlying context that defines existing business processes (e.g. staff trust ethic) and then derive the current business rules for the organisation.
- Examine this context and seek corporate change that enables effectiveness and efficiency.
- Eliminate those business processes now no longer needed.
- Automate the remaining processes in the context of an overarching Business Model.
- Develop a Data Model to encapsulate all business processes.

Objective: to integrate all existing information systems into a logically single Corporate Data Repository (CDR).
Tactics:

- Develop and agree a CCLRC-wide terminology defining the information terms used in the Business Model by first developing a suitable Corporate Data Model (CDM). The CDM must be flexible and robust against change.
- During the examination of all existing paper and electronic forms and the identification of context (data and process) issues, discuss, identify and define one set of terms and use these terms in the definition of the CDM. In this way the context change defined in these terms (such as a change in the staff trust ethic) applies across all related processes and their associated forms.

In 2001 the Business Process Sub Group (BPSG) was set up to progress this strategy. All current business forms were collected, grouped into process categories and analysed. An initial CDM was constructed based on the three CERIF primary entities of Person, Organisation Unit and Project. Our entities were in fact called *Person*, *Organisational Unit* and *Activity* and were initially derived independently of CERIF. As with CERIF, all primary relationships had link entities which were dated to record their changing valid status (the 'when'). The Person entity modelled anyone with a business relationship with CCLRC (the 'who'); Organisation Unit reflec-

ted the CCLRC organisational structure (the 'context') and Activity reflected business processes of all kinds (the 'what'). Therefore, this model captured change by the "what/when/who/context".

Furthermore, it was felt that a process data model should fit into the CDM, thereby tracking process changes automatically and by instancing these processes we could record the execution of a process by the same dated relationship mechanism and record "who did what when and in what context".

It was clear from our initial exploration that the work was feasible and that we needed to do it for real in some limited pilot study to provide a cost-benefit analysis for a business case to undertake a comprehensive implementation. Funding for the pilot is now approved, with the main tasks as:

- Collecting and recording all forms and terms and analysing the data.
- Identifying and recording on the Web two or three pilot business processes and their associated business rules in the context of a light examination of all processes.
- Synthesise issues; propose new enabling business rule revisions.
- Implementing the prototype CDM - resolving detailed semantic aspects with a view to showing that CERIF can indeed be used in this context – and delivering an initial, partially populated CDR.

3 CDR Requirements

When CDR development at CCLRC moved into its pilot phase, a further examination of the functional requirements was made in preparation for a first-cut implementation of a CDM. The feasibility study outlined in the previous section defined the benefits to the organisation as a whole. In this section we shall look at the CDR requirements in terms of existing systems and project requirements.
A central objective of the CDR is to develop a single authoritative source of shared corporate memory that cuts across site, department and temporal boundaries. Far from being a passive recipient of corporate data, the CDR has a pivotal role in the planned Corporate Systems (CSSB, 2003). For example, it serves as a data exchange hub where relevant information on organisational structure, resources and operations are validated and exchanged with function-specific information systems such as those outlined in Section 5.

In the previous section we identified the CCLRC feasibility study requirements to facilitate business process re-engineering using workflow technologies by:

- Identifying corporate business processes.
- Establishing a CCLRC-wide common vocabulary for the business processes.

- Formalising the business rules that underpin these business processes.
- Capturing the terms, business rules, processes and their inter-relationships in a CDM.

In addition to these we also have the following corporate requirements:

- Helping compliance with legislation in data privacy and freedom of information by being the single authoritative source of both current and *historical* corporate information.
- The integration of existing corporate systems by acting as a data hub for various specialised systems. Since CCLRC is a complex, multi-site organisation, those existing corporate information systems at its *Daresbury*, *Rutherford Appleton* and *Chilbolton* sites need careful examination.
- Facilitating the development of new corporate information systems by ensuring that the CDR can supply accurate, up-to-date and comprehensive corporate information upon demand. Current CCLRC projects include the *Information Portal*, *e-Library* and *e-Record Management Systems*.

In addition to these business requirements, CCLRC as a *research* institution recognises its relationship to CRIS community at large and particularly euroCRIS and CERIF. Our CDM is developed with reference to the CERIF-2002 standard (euroCRIS, 2003) and our expressed goal is to conform to that standard and feed back our experiences to the CERIF task group.

An obvious problem, however, when using the CERIF-2002 standard as a starting point for the CCLRC corporate data model is that CERIF was designed with *current research* information systems (CRISs) in mind and not *corporate* information systems. A large number of entities and relationships are missing from CERIF that are required for modelling a complex organisation such as CCLRC. For example, to model our human resources adequately would require a data model similar to that of *Oracle Human Resources*. Similarly, to model the company accounting system would require a complex data model similar to that of *Oracle Financials*. Other areas not covered by CERIF-2002 include health & safety, site security, car parking, facility & services management, process manufacturing, business processes, corporate procedures, contract management, laboratory testing, computer support & development, property management, material requirements planning, project procurement & management, travel, expenses, taxi bookings and documentation – to name just a few obvious examples. It is hardly surprising that some eyebrows were raised when at first it was proposed to use the CERIF standard to model the CCLRC enterprise.

However, the CERIF data model is a simple yet flexible structure. For our purposes it needs extending but the feeling is that the data model is flexible enough to adjust to changes in our corporate requirements. At this stage it is *not* proposed to model

the financial or human resource subsystems already covered by other software packages. For the pilot study we propose to implement only two key business processes which will have a general impact on the organisation as a whole. These will demonstrate the proof of concept sufficiently to attract further funding for extending the CDM and developing an integrated CDR that will benefit the whole of CCLRC.

4 CDM Implementation

In this section we shall examine the changes made to the CERIF-2002 data model to meet the needs of two chosen CCLRC business processes (*Reporting Sick Absence* and another process currently being selected). We shall start with a discussion of the generalisation of the CERIF data model, followed by discussions on our use of entity histories, transactions times, entity subtypes, relationship subroles and relationship hierarchies. Our development environment includes the ER*win* 4.1 data modeller, *SqlEdit* 1.6 (White, 2003), ODBC drivers for Oracle 9i, Microsoft SQL Server 2000 & IBM DB2 8.1, running on Microsoft Windows NT/XP. The workflow management system for the pilot study uses a open-source workflow engine (Zope, 2003) running on Linux. A proprietary workflow system will been chosen to implement the other process.

4.1 CERIF Generalisation

A general system comprises a set of interrelated parts operating in a coordinated way to perform one or several functions. Such a system is "*more than the sum of its parts*" precisely because these relationships constrain the parts in new ways, allowing different degrees of freedom and a more complex behaviour than the parts themselves. These parts together with their inter-relationships manifest a new *system* behaviour with enhanced functionalities.

In the CDM we subdivide the system into three main functional areas, as illustrated in figure 4.1. These subsystems are called *Organisation*, *Operation* and *Resource*. The *Organisation* subsystem records the structure of the system parts, the *Operation* subsystem records the operations and activities of the system parts and the *Resource* subsystem records information about the productive resources that perform the operations of the system. Note that each subsystem records an aspect of the modelled external *Universe*. We can use this general system architecture to motivate a generic data model based on these three subsystems.

Figure 4.2 shows the top-level entities of a General System Data Model based on the General System Architecture of figure 4.1. This data model is used at the top-

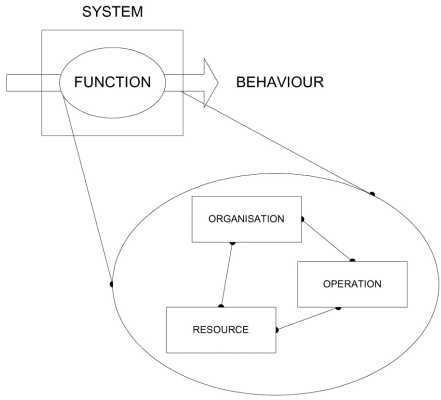

Figure 4.1 – General System Architecture

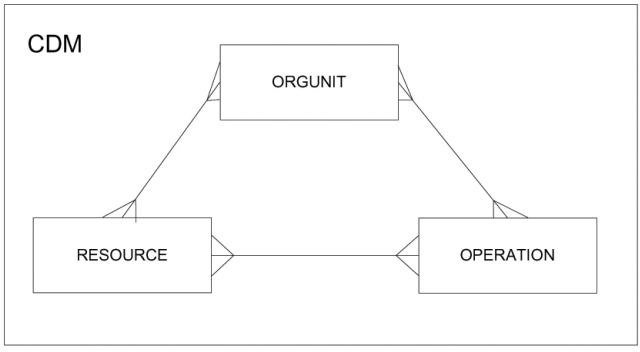

Figure 4.2 – General System Data Model

level of the CCLRC CDM. There is a striking resemblance between the CERIF Data Model and our General System Data Model. Indeed, CERIF can be viewed as a specialisation of that data model, as demonstrated in figure 4.3 where *OrgUnit* maps to *OrgUnit*, *Resource* to *People* and *Operation* to *Project*.

With such clearly delineated subsystems it is much easier to map new entities to the functional areas of the CDM in figure 4.3

For example, a *company car* would be a Resource owned by the Organisation. A *service* would be an Operation performed for the Organisation by a Resource. The term *Resource* now includes the sub-types *human* Resource (i.e. person) and *physica*l Resource (e.g. company car, building).

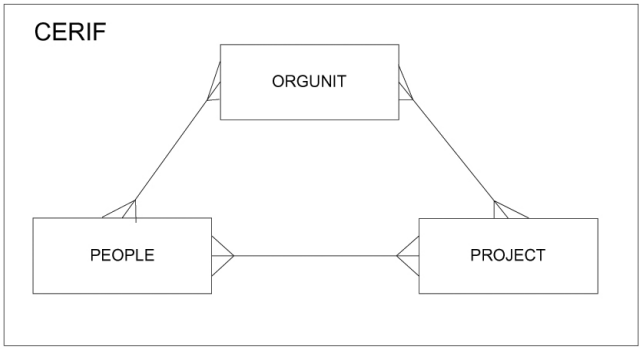

Figure 4.3 – CERIF Data Model

Our approach to modelling the CCLRC CDM means there is an upwards compatibility between CERIF-2002 and our extended CDM. Or, to put it another way, CERIF-2002 is a subset of the CDM. Note that for the purposes of this discussion we have ignored the recursive relationships on each entity, which we have in fact implemented.

4.2 Entity History

CERIF-2002 makes the assumption that only *relationships* have a dynamic history. For this reason only *relationships* having a valid time *start* and *end* and are recorded in the CERIF data model. For example, John Smith (person) worked on project *Tornado* between 12 January 2000 (*start*) and 26 March 2003 (*end*). Such relationships are what CERIF was primarily designed to record.
However, in *corporate* data models we find that *entities* are also dynamic and worthy of having a history. For example, on the 1st January 1999 Jan Kowalski changed his name to John Smith. CERIF is not able to record this fact because it assumes that entity names do not change. Similarly, if Jan Kowalski changed his gender from male to female and changed his name to Janina Kowalska then, once again, CERIF would be unable to record this fact in the database. Whilst this may not seem a serious omission in a CRIS database, it *is* a serious deficiency in a corporate database where pensions *must* be calculated accurately. For this reason we decided to include a valid *start-date* and *end-date* for *entities* as well as relationships in the CDM. In effect we have extended the CERIF-2002 data model and implemented a *temporal database* with tuple timestamping (Skjellaug, 1997). Each *row* in every CDM table is associated with a *valid time*.

4.3 Transaction Time

The CDM has a separate *Transaction* table which stores the *transaction time*. Whereas *valid time* records when we believed a proposition was true in the enterprise, *transaction time* records when we stored that true proposition in the database (Date et al., 2003). For example, John Smith may have changed his name on 1st January 1999 (valid time *start-date*) but this fact may have only been recorded in the database on 12th January 2000 (transaction time *start-date*).

We timestamp each tuple with a valid time and transaction time. Furthermore, each tuple is uniquely identified by a *Database Identifier*, *Internal Identifier* and *External Identifier*. The *Database Identifier* uniquely identifies a Conceptual level database schema (see section 4.5 *ANSI/SPARC Architecture*). The *Internal Identifier* is a surrogate key (Date, 2004) for the entire database schema and is updateable only by the database system. It is invisible to the External user application. The *External Identifier* is a user key that is updateable by the user application. Each table primary key comprises all three identifiers, ensuring that each tuple is unique across database schema.

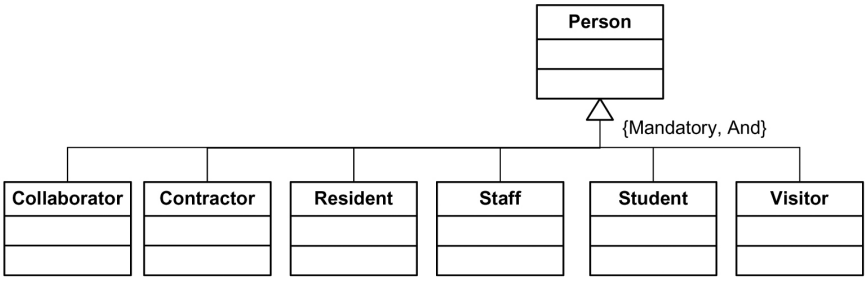

Figure 4.4 – The CDM Person subtables (partial set)

4.4 Entity Subtables

In the CDM we have extended the CERIF entities to include *subtables*. For example, the *Person* entity may have a variety of subtables to facilitate the representation of multiple roles, such as *Student* and *Staff* (see figure 4.4). Note that a *Person* entity may be simultaneously a member of Staff *and* a Student. Given the importance of a person's role in determining business process flow and in capturing corporate memory (the '*who did what when*' aspect), implementing a *Person* hierarchy by creating subtables helps enforce more specific business rules (Barker, 1989). We are aware of the stance of some authors on the issue of subtables (Date and Darwen, 1998) however we prefer to think of *subtables* as relation *subschema*.

Should it prove necessary, we will extend our use of *subtables* to the link entities themselves. For example, the *Person_OrgUnit* supertable might have *Staff_OrgUnit* and *Student_OrgUnit* as subtables, since all *Staff_OrgUnit* and *Student_OrgUnit* relationships must be a subset of *Person_OrgUnit* relationships.

4.5 ANSI/SPARC Architecture

The *ANSI/SPARC Architecture* so widely discussed in text books on database theory (Date, 2004) is seldom adhered to in practice and hence the major benefits of *data independence* are lost in many database systems. To be fair, the main reason for this has been the lack of updateable views on multiple tables, without which it is difficult to see how the three level architecture (External, Conceptual and Internal levels) can be implemented.

After splitting the Conceptual level into two further sublevels - the *conceptual* and *logical* sublevels - we implemented the Conceptual-*logical* level with tables and the External and Conceptual-*conceptual* level with updateable views using a non-standard SQL statement CREATE TRIGGER … INSTEAD OF supported by Oracle (also by Microsoft SQL Server and IBM DB2. An equivalent CREATE RULE statement exists in Ingres and PostgreSQL).

4.6 Data Distribution

An examination of current CCLRC databases shows that much data is copied from one database to another, often via an ODBC application written in Microsoft Access, which effectively acts as a database interface, mapping data from one data model to another. This approach will probably continue for as long as CCLRC has a heterogeneous databases environment. However, there is no reason why such constraints should be placed on the CDR. With an eye to the future, it was anticipated that the CDR may be distributed across sites. By this we mean that a copy may run at RAL and two further copies at DL and CF (if required). This would allow local CDR databases to run in parallel on different sites, however, they would need to share data periodically e.g. overnight downloads. Care would be taken to ensure that the data is exchanged frequently enough to give the user the impression that there was a single CDR but without overloading the inter-site network. Data replication is one way of achieving this data transparency across sites. The system has been designed from the outset to assign unique identifiers to all table rows within the enterprise through a *database identifier*, forming part of the primary key of each table row. Each *database identifier* is unique across all CCLRC sites. It is therefore possible to mix rows between databases without fear of creating identical primary key values across sites.

5 CDR Architecture

How is CCLRC going to use the CDR as a corporate data hub connecting other specialised databases and applications?

There are several corporate databases holding information that is central to performing many of the business processes required in CCLRC. Many of these processes require access to the similar information about people, projects, services, organisational roles, organisation structures and financial information. In order to avoid the many pitfalls of entering information many times, each for a specific purpose, it is essential to have some mechanism for sharing information. One of our goals is to input information only once, as close as possible to it's source. This information is then shared by business processes that need it, in a secure and managed way, to ensure that privacy and other issues are upheld.

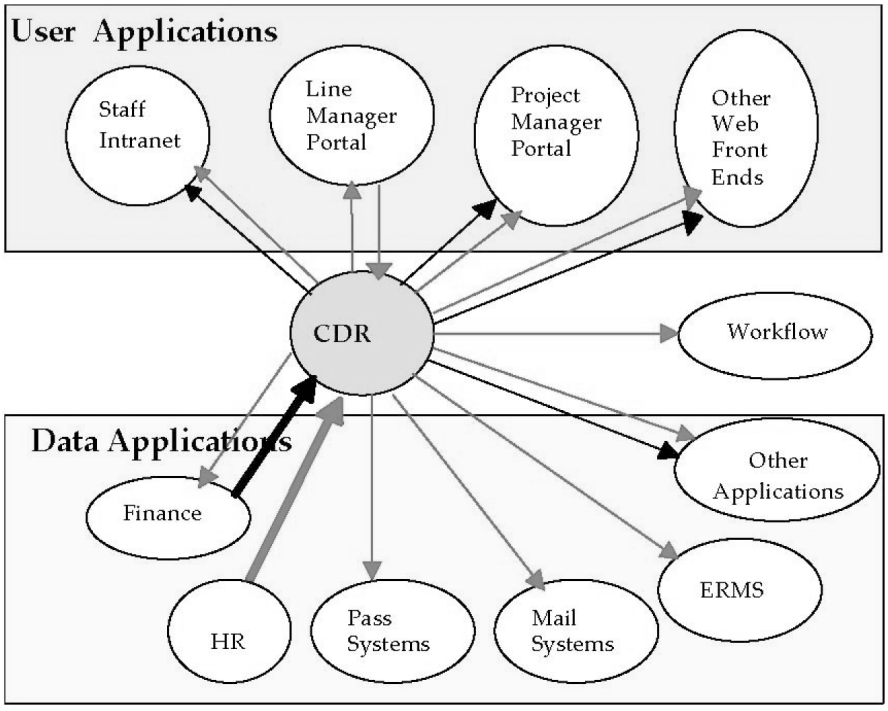

Figure 5.1 – CDR Architecture

The CDR acts as a hub through which information can flow between databases, data applications and user applications (see figure 5.1). It acts as a conduit to the *Information Portal*, staff Intranet, and other user-focussed applications such as form filing, smart-card swiping, etc. The aim of the CDR is to be a "one-stop shop" for information retrieval and information storage. Specialised input interfaces closely associated with the *local* data storage can upload data into the CDR as required. Normally these specialised systems only export a sub-set of their local data store since usually only some of this data will be of *general* interest. On the other hand, specialised data applications are able to retrieve information from the CDR rather than having to store a local copy of the data. To implement CCLRC business processes, our workflow management engine connects to the CDR database and retrieves information about people, projects and roles.

In order for the CDR to provide the information as and when required, interfaces are implemented to provide the information in a timely manner, supplying it either as a bulk upload/download or as a "one-shot" request. These interfaces, utilising ODBC or JDBC connections to the database, ensure that the integrity of the information is maintained. Checks are made each time some information is presented to the CDR and alarms are raised as soon as a problem occurs.

From a security viewpoint, bad data is more of a problem to the organisation than no data at all. Interfaces with the CDR are secure and managed so that information is only provided to those applications and users permitted to see it and use it. This is achieved using LDAP and Microsoft's Active Directory.

Figure 5.1 shows some of the applications using the CDR together with the major data flows. The primary information sources to the CDR are from our Human Resource (HR) and Finance subsystems and there are also flows to provide project, role and resource information to other applications. In the diagram, arrows indicate the direction of information flow; the width of the stem indicates the volume of information and the colour signifies the information source (black for Finance; red (or grey if viewed in monochrome) for HR).

Finally, the *Information Portal* project currently being developed at CCLRC is creating an user-configurable Web interface to view corporate Web pages but using the CDR to determine user privileges to CCLRC corporate information. In future it is hoped to create a Semantic Portal, using ontologies being developed in-house to permit more "intelligent" user queries and a rule engine to aid the definition of business rules (Date, 2000).

6 Acknowledgements

We would like to thank our colleague at CCLRC, Anne Shrimpton, for supplying the CDR Architecture diagram.

7 References

Barker, R. (1989): *CASE*METHOD Entity Relationship Modelling*. Harlow:Pearson Education
Ltd.
BSSG (2001): *BSSG Business Systems Strategy (final draft)*. A CCLRC internal CM Boardpaper.
CSSB (2003): *CSSB Corporate Systems Strategy (first draft)*. A CCLRC internal CCSB Committee paper.
Daniels, T. (2002): *Database Rationalisation (2).* An CCLRC internal CSSB Committee paper. Date, C. J. (2004): *An Introduction to Database Systems* (Eighth Edition). Pearson Education Inc., Addison-Wesley, USA.
Date, C. J. (2000): *What Not How: The Business Rules Approach to Application Development*.
Addison-Wesley, Pearson Education Inc., New Jersey, USA.
Date, C. J. and Darwen, H. (1998). *Foundations for Object/Relational Databases: The Third*
Manifesto. Addison-Wesley Publishing Company Inc., USA.
Date, C. J., Darwen, H. and Lorentzos, N. A. (2003): *Temporal Data and the Relational Model*.
Elsevier Science (USA), Morgan Kaufmann Publishers, San Francisco, USA.
euroCRIS (2003): *CERIF*. www.eurocris.org.
Skjellaug, B. (1997): *Temporal Data: Time and Relational Databases*. University of Oslo.
White, J. (2003): *SqlEdit 1.6*. www.planetepoch.com.
Zope (2003): *Zope 2.5*. www.zope.com

8 Contact Information

Trudy Hall
CCLRC – RAL
BITD
Fermi Road Chilton
Didcot
Oxfordshire OX11 0QX

e-mail: Trudy.Hall@rl.ac.uk