PhUSE 2006

Paper AD09

Contours of a Drug Development Data Warehouse

Paul Vervuren, Organon, Oss, The Netherlands
Frank Dietvorst, PW Consulting, Culemborg, The Netherlands

ABSTRACT
Organon is conducting a data warehouse project that aims to make drug development data more easily and widely available across the R&D organization. Main goal is to enable study data exploration across compounds. Initially, the project will be limited to disclosing clinical data. This paper presents our considerations and solutions for the various warehouse and toolset components, including the data warehouse architecture and software/hardware design. In addition, it discusses the role of CDISC and the possibilities of aligning the data warehouse with the clinical data process

INTRODUCTION
The clinical data process is not designed for cross-compound data exploration or to satisfy information needs in general outside the scope of study reporting and submission. In particular, the conventional clinical data process has the following limitations:

- Direct access to (analysis) data is mainly restricted to programmers and statisticians;
- Data is not integrated above the compound level;
- Keystroke programming is common, even for basic analysis and reporting;
- Study-level information (e.g. clinical phase, trial duration, type of control), which is often involved in querying, is not available in databases.

Led by the desire to offset these limitations and encouraged by external developments, in particular CDISC and the FDA's Critical Path Initiative, Organon has initiated a data warehouse project (Drug Development Data Warehouse: D3W). The ultimate aim of D3W is to enable the exploration of data gathered in the drug development process (from drug discovery to post-marketing) to help accelerating and optimizing the product development pathway, and to service a variety of more general information needs pertaining to drug development data (e.g. trial overviews for Periodic Safety Update Reports). In order to make these ambitions manageable the initial scope of the project is narrower, concentrating on data exploration for hypothesis-generating purposes using clinical data.

Earlier phases of the project, which started in 2004, laid a solid foundation for the project (see Vervuren 2005). Now, having made the principle choices for hardware and software and being in the process of designing the data architecture, the contours of the data warehouse are emerging. This paper presents our considerations and solutions for the D3W data warehouse and its user tools, including the data warehouse architecture and software/hardware design. In addition, we will discuss the role of CDISC and the possibilities of aligning the data warehouse with the clinical data process.

THE ROLE OF CDISC
Data standards are essential to data warehousing. CDISC provides clinical data standards that have been widely acknowledged by the pharmaceutical industry. The FDA endorsed CDISC's SDTM model as a submission standard in 2004. SDTM is the model for study tabulation data, i.e. 'individual observations for a subject that comprise the essential data collected in a clinical trial' (SDTMIG V3.1.1). For D3W we adopted the SDTM nomenclature, class and domain definitions, its directives regarding changing domains or designing new ones, and its identifier and timing definitions.

CDISC models, as interchange standards, are optimized to enable clear and unambiguous communication on data; they do not present ready-made data models for data warehousing. The design of our data warehouse is based on a partly restructured version of SDTM, optimized for data integration

and analysis (see next section). Moreover, to support Organon-specific demands extensions of the SDTM domains (SDTM+), comparable to Accovion's "SDTM comfort" (Arnold & Plank 2005), and additional sets of variables will be required. This may concern information from internal operational processes or data required for cross-compound querying and analysis, e.g. compound characteristics and classifications. The added variables will follow SDTM principles for variable definitions. Thus, the CDISC SDTM domain variables and Organon SDTM extensions form a new, internal superset of SDS Variables (SDS+). Although the CDISC SDTM and Organon-specific SDTM definitions are based on the same rules, they will partly have their own dynamics. CDISC SDTM variable definitions will keep pace with the official versions, while Organon SDTM variables are internally managed. Although the current emphasis lies with SDTM, incorporating basic usage of ADaM is essential for analysis data. The D3W approach to ADaM is similar to SDTM, resulting in ADaM+.

The SDTM model allows a certain degree of freedom and some elements are sponsor-defined. The question is how to deal with this variability. For individual submissions, any variation within the boundaries of the SDTM implementation guideline may be acceptable. However, to facilitate full data integration certain variations may be unwanted. For instance, the addition of derived records (e.g. for averages; see SDTMIG V3.1.1, p. 148) in the LB domain. Solutions or constraints optimal for a multiple compound data warehouse may not be in the field of vision of someone preparing data for a single submission. We therefore suggest that the superset of SDS variables as well as the interpretation of CDISC guidelines should be centrally defined and managed, serving both data warehouse purposes and the operational process.

Supporting the submission process is outside the current scope of the D3W project. However, with the selection of SDTM as a data standard for our clinical data warehouse creating submission data becomes within reach. For all warehouse data based on Organon design principles mapping rules can be defined that describe how to generate submission data, for example controlling which extensions to SDTM must be loaded into a supplemental qualifiers data set or serve as metadata for define.xml.

DATA WAREHOUSE ARCHITECTURE

As subject data flows through the operational clinical trial process the data models vary, suiting the needs of each process step. Clinical data management systems like Oracle Clinical are typically based on an entity-attribute-value (EAV) model, with the advantage of providing low-maintenance flexible storage. Any new clinical parameter simply represents a new row in the EAV database (e.g. see Brandt et al. 2002 for a discussion of the EAV model in the context of a clinical database). The alternative is a column-oriented or representational model, which is more suitable for analytical purposes. Hence, this model is seen in the late stages of the process, supporting analysis and reporting. Usually, as is the case with Organon, data used for such purposes is made analysis-ready through the addition of 'redundant' variables (treatment code, analysis-group flags, some demographics variables, site of investigation, investigator).

The requirement of storing multi-trial data pleas for an EAV approach for clinical data warehouses (Hughes, 2004). Developing a standard representational model for clinical data seems an impossible mission. CDISC is good news in this respect, and provides a basis for representational modeling of clinical data warehouses.

The modeling approach adopted for the clinical trial data within D3W is a hybrid one, with a representational model for commonly captured data and an EAV model as a generic storage solution for data captured specifically within a study. Figure 1 shows a schema of the proposed D3W warehouse architecture, with domain-specific storage tables (DWH domain units), an interface layer (the software that enables automated data loading and extraction), a metadata layer containing the required technical and business information to support the entire process (including end-user tools), and a lookup layer to support querying. The advantages of this approach are:

- Standard data are in a representational form, thus:
  - suitable for query and analysis
  - largely self-explaining, and hence
  - transparent and controllable
- Capable of loading all clinical data; CDISC compliant and non-compliant
- Capable of creating CDISC compliant data (e.g. for submission)

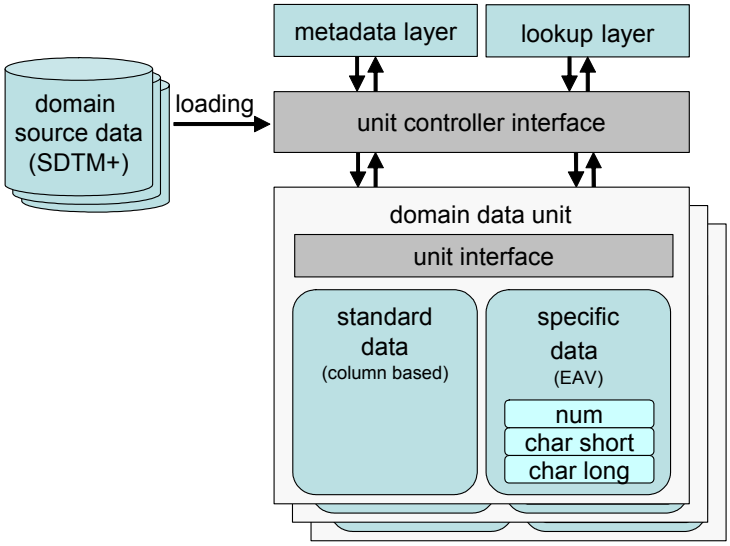

Figure 1. Schematic representation of the proposed D3W data warehouse architecture. The distribution of standard and study-specific data is domain specific. For efficiency reasons, specific data contains separate EAV structures for different data types (num, char short, char long).

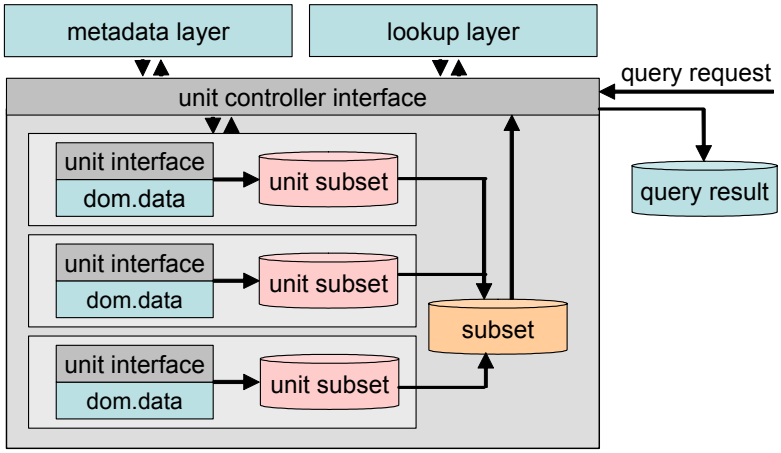

Figure 2. Schematic representation of multiple domain units combined to generate a query result.

Where domain units represent a specific data collection, for example an SDTM+ domain, queries mostly require a broader scope (fig. 2). It is not an option to have the user collect and subsequently combine all data unit subsets, as this will compromise user friendliness, and, more importantly, introduce major data reliability risks. To automate this process, all queries are managed by the unit controller interface. This interface can analyze the query and generate sub-queries for all involved data units. When all data unit subsets are available, the controller interface combines the subsets. The original query is applied to this subset, and finally the query result is returned to the user.

SOFTWARE AND HARDWARE DESIGN

To support quality control and production stability, a separated development environment and production environment have been planned. The idea is to start-off with a fully functional development environment (with access limited to test users), which will serve as a production prototype. Based on benchmark and test results the final production environment will be defined. Deployment of the final production environment depends on:

- Standardized clinical data, including study-level information captured in database format
- Documentation and validation of essential processes (e.g. data transformations)

- Data confidentiality arrangements
- Staffing (data administration, technical and functional management)
- Alignment with the clinical trial process

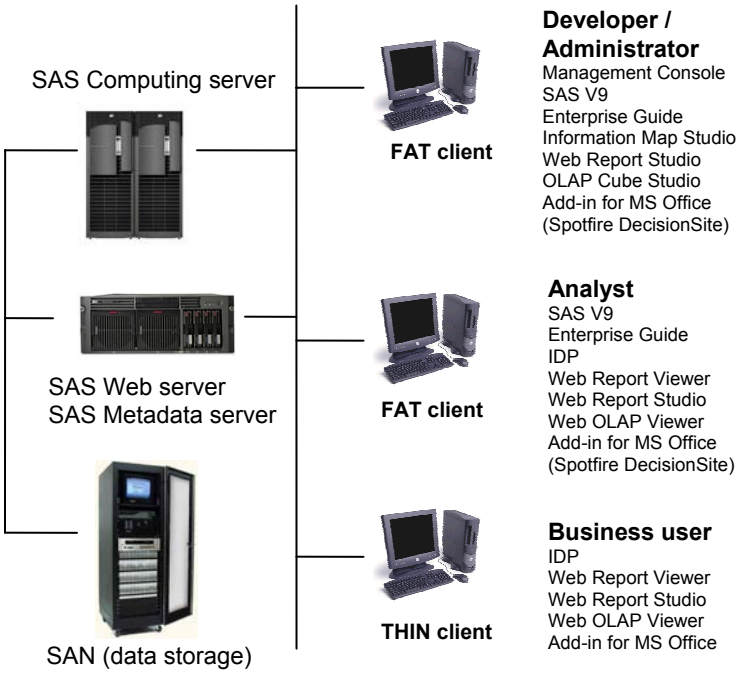

Figure 3. Overview of the proposed hardware and software design of D3W.

Organon has the intention to use the SAS®9 architecture as the basis for the D3W data warehouse and front-ends (fig. 3). Decisive criteria in the selection process were:

- Complete and integrated software suite
- Central metadata
- Open architecture
- Scaleable
- Involved in pharmaceutical industry and active participant of CDISC
- Internal SAS expertise available
- SAS is an accepted standard for statistical analysis

For storage native SAS file formats will be used. The well-developed connectivity of SAS enables an easy switch to a different format if desirable.

SAS Stored Processes and Information Maps will be used as interface layer between the data repository and the front-end tools. This ensures the use of identical routines for accessing and querying the data repository for all front-ends, fat as well as thin client. SAS Enterprise Guide, with the full data processing and analytical capabilities of SAS, serves as a versatile user-friendly tool for analysts. Complementary to SAS Enterprise Guide, we envision a possible role for Spotfire DecisionSite, which strength is interactive data visualization and filtering. Organon already deploys Spotfire for interactive analytics in support of its basic research. Business users are offered controlled data access by means of web reports and stored-process applications via SAS Information Delivery Portal.

We currently focus on locked data that is integrated at a compound level, but is not SDTM standardized. Custom mapping to SDTM+ will be needed to allow the loading of this data in our data warehouse. Earlier we considered deployment of SAS Data Integration Studio (previously ETL Studio) for the ETL process (data extraction and transformation and loading into the warehouse). Our current position is that optimal deployment of SAS DI Studio requires the clinical data process to supply SDTM+ standardized data.

PROCESS ALIGNMENT

Usually data warehouses have a position at the end of or parallel to operational process chains, consolidating operational data for strategic purposes. The D3W data warehouse, at least regarding the clinical domain, has the potential to become fully aligned, if not joined, with the study reporting and submission process. This is the result of two favorable conditions:

- CDISC provides a common standard
- The stages of the process chain that provide analysis-ready data are already conducting extraction and transformation processes (yet rather custom for each compound); they may form the starting point for ETL processes that build the data warehouse

The benefits of this full alignment would be:

- Single source of data
- Single effort for preparing data and validation
- Modernization of the clinical data process
- Shared use and management of (D3W) facilities

Since existing processes are just starting to prepare for CDISC, our clinical data warehouse will be ahead of time in terms of data standardization. Alignment of the data warehouse with the operational clinical data process will be a gradual development involving the implementation of CDISC in this process. Figure 4 depicts a possible scenario for this alignment, which will require or invoke operational data process optimizations and changes.

A. Currently the operational data process (above) follows internal standards (blue color). The warehouse data (below) follows CDISC standards (orange color). The data warehouse will initially load locked data, which will need to be converted to SDTM+ and ADaM+ before loading. In the absence of a suitable storage format, most data elements for the SDTM Trial Summary domain (consisting of protocol elements) need to be reentered and are added to the trial data at the end of the chain.
B. Here, CDISC is implemented at the late stage of the operational data flow (analysis and reporting), delivering Organon extended versions of SDTM and ADaM: SDTM+/ADaM+. This standardization will greatly facilitate integration and enable the loading of this data into the data warehouse. In this situation, the creation of SDTM+/ADaM+ can be incorporated in the analysis and reporting process or take place afterwards. Protocol elements are preferably recorded early in the process, e.g. as part of the protocol writing process, and stored in the CDMS (for instance). (In this phase the data process is already more prepared for CDISC compliant submissions, but the late-stage conversion to SDTM/ADaM will still impede the analysis and reporting step.)
C. This presents a situation where CDISC standards have been implemented in the entire process chain (‘end-to-end’), making the clinical data process more transparent and easier to manage. Removing the need for conversion to SDTM+/ADaM+ will relieve the analysis and reporting step. The use of a common data standard will allow shortcuts. For instance, ongoing study data could be directly loaded into the data warehouse, enabling data review via analytical tools of the D3W environment like SAS Enterprise Guide or Spotfire.
D. This step speculates on a further integration between the data warehouse environment and the operational data process. In this situation the data warehouse facilities support both new applications (e.g. cross-compound exploration) as well as existing operational processes (data review, analysis and reporting, submission).The SDTM data in the CDMS has matured from ‘pre-SDTM+’ to ‘SDTM+’, allowing direct loading into the warehouse. This would further relieve the analysis and reporting step allowing it to fully focus on ADaM+.

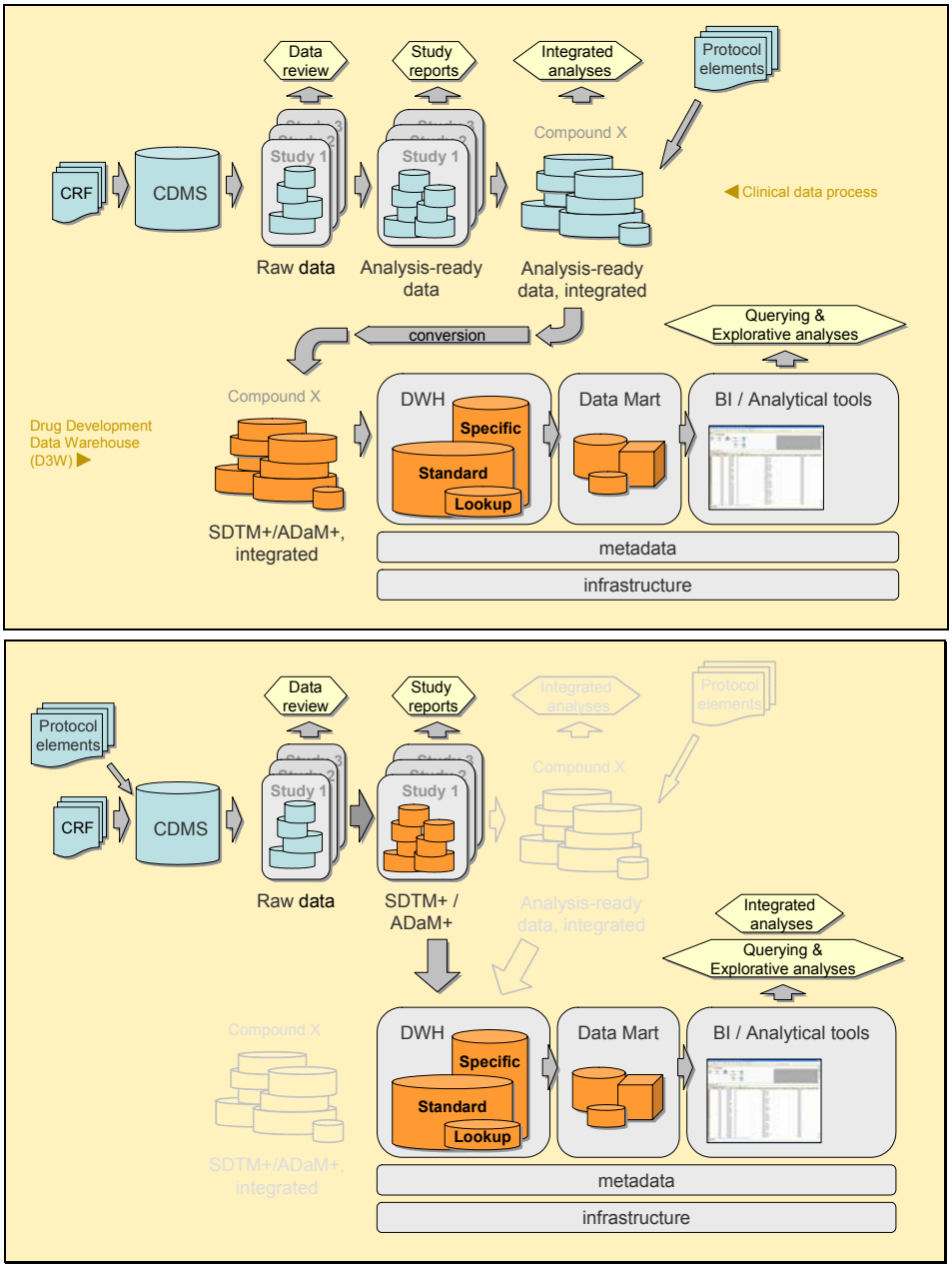

Figure 4. Hypothetical steps for D3W implementation, involving integration with the clinical data process. Components in orange use CDISC standards. See text for explanation.

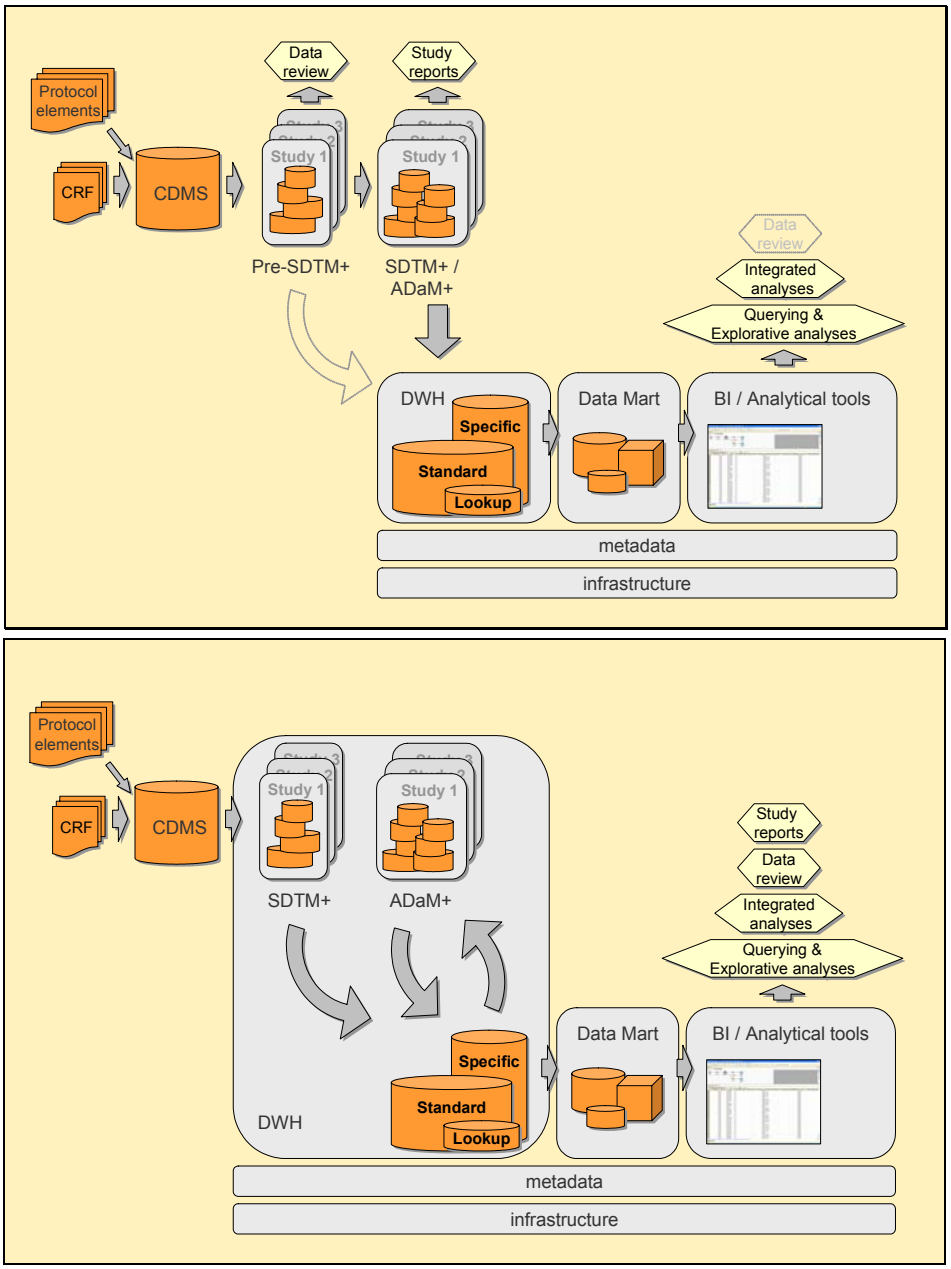

Figure 4. Hypothetical steps for D3W implementation, continued.

The above does not draft a planned route for implementation but pictures a hypothetical scenario that intends to stimulate discussion on implementation options and desired or required process improvements. It emphasizes the importance of CDISC for our data warehouse project, both within our current scope as well as for future options.

CONCLUSION

This paper presented our considerations and solutions for the D3W data warehouse and its user tools. In summary:

- CDISC is essential to D3W in providing the basis for the warehouse data model, and representing the future data standard for our source data. We propose extensions to the CDISC models (SDTM+/ADaM+) for internal use, in order to satisfy Organon's business requirements. The definition

and management of the new internal superset of SDS variables thus formed must become a shared task of operational and warehouse data managers. Structural modifications to SDTM/ADaM are planned in order to optimize the data for usage in D3W.

- As part of the D3W data warehouse architecture a hybrid model (partly representational, partly EAV) has been designed for clinical data, for which CDISC SDTM was used as a starting point. Design considerations were: capability of loading all clinical data, suitability for querying and analysis, and support for CDISC compliant data submission.
- Principle decisions for hardware and software have been made and implementation is being prepared. Aim is to create a prototype production environment, based on SAS9 software, that provides facilities to different types of users: developers, analysts and business users. SAS Enterprise Guide is the proposed tool for analysts, possibly complemented with Spotfire. Business users will be offered web reports and stored-process applications via SAS Information Delivery Portal.
- Clinical data process changes would improve the alignment of D3W and the clinical data process. CDISC and other developments like the FDA's Critical Path initiative are important drivers for our project and create momentum for process changes.

Organon's D3W project has progressed to a stage in which the contours of the data warehouse are emerging. Next steps will witness the realization of the drug development data warehouse, starting with the inclusion of clinical data. Anticipated activities in the near future are: (1) detailed design of SDTM+/ADaM+, (2) hardware and software implementation, (3) testing and piloting the data warehouse architecture, (4) developing a framework for managing D3W data and facilities.

REFERENCES

Arnold E. & U. Plank (2005). "Customer oriented CDISC implementation", *Proceedings of the First Conference of the Pharmaceutical Users Software Exchange (PhUSE)*.

Brandt C.A., R. Morse, K. Matthews, K. Sun, A.M. Deshpande, R. Gadagkar, D.B. Cohen, P.L. Miller & P.M. Nadkarni (2002). "Metadata-driven creation of data marts from an EAV-modeled clinical research database", *International Journal of Medical Informatics* 65: 225-241.

Hughes. R. (2004). "Optimal Data Architecture for Clinical Data Warehouses", DM Review, Nov. 2004. (www.dmreview.com/article_sub.cfm?articleID=1012400).

STDMIG V3.1.1 (2005). "Study Data Tabulation Model Implementation Guide: Human Clinical Trials", Document version 3.1.1. Submission Data Standards Version 1.1.

Vervuren P. (2005). "Visions of a Drug Development Data Warehouse", *Proceedings of the First Conference of the Pharmaceutical Users Software Exchange (PhUSE)*.

ACKNOWLEDGMENTS

We are greatly indebted to Goran Cizmedic for his vital contribution to the development of the data architecture. We would like to thank Kit Roes and Egbert Biesheuvel for their guidance and for reviewing the manuscript.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. You may contact us at:

Paul Vervuren
Biometrics/Clinical Information department
NV Organon
PO Box 20
5340BH Oss
Work Phone: +31 412 663921
Email: Paul.Vervuren@organon.com
Web: www.organon.com

Frank Dietvorst
Sr. Consultant
PW Consulting
PO Box 373
4100AJ Culemborg
Work Phone: +31 6 1092 0122
Email: Frank.Dietvorst@tiscali.nl
Web: www.pwcons.com

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.