

# HDI: Research Software to Commercial Product

D. P. Hansen, C. Daly, K. Harrap, J. Jacquet, M. O'Dwyer, C. Pang and J. Ryan-Brown.  
*e-Health Research Centre, CSIRO ICT Centre, Brisbane*  
*david.hansen@csiro.au*

## Abstract

*Making full and effective use of health related data sources is seen as a key way for Australia to improve the efficiency of its health system. Major technical and legal questions exist concerning data integration, data quality, data security and privacy in health data usage. In this paper we discuss the Health Data Integration™ (HDI™) solution which addresses these issues while providing a range of useful functionality. The e-Health Research Centre is responsible for the further development and release of HDI™ and is working with Queensland Health to deploy the software in practice.*

## 1. Introduction

Medical and related data is currently recorded in a variety of places, for a variety of reasons and in a variety of formats. Most data related to a health event is recorded to ensure that the provider is able to track costs, rather than as a way of recording clinical treatment regimes and the efficacy of that treatment. This makes the analysis of this data difficult for purposes such as clinical improvement, health service provision and health policy development. Bringing this data together for meaningful analyses is hindered by these technical aspects. In addition the linking of data sets raises considerable ethical and privacy issues for the patient and the data custodian.

The Health Data Integration (HDI™) software solution began as a research project within CSIRO [1] for providing this functionality to researchers, clinicians and policy researchers. The software is being further developed at the e-Health Research Centre (EHRC) [2], a \$15 million joint venture between CSIRO and the Queensland Government. The aim of this further development is to transition the software from a research project to a product which will be used within Queensland Health as a valuable tool for clinical improvement and policy development. Another aim of the EHRC is to

commercialize the HDI™ software solution for other users.

The requirements of developing software for research purposes to solve a particular problem are often very different to those of developing a software product. For instance, research software need not meet the same standards of robustness and maintainability as would be expected in commercial software.

HDI™ offers novel solutions for linking patient data while preserving patient privacy. This is a cornerstone of HDI™ and the result of considerable research work thus far [1, 3]. However, HDI™ will only be accessible to a wide range of users if this functionality is offered as part of a complete package. The complete package must allow data custodians to add their data to an integrated data network, and must provide end-users with the ability to analyze and produce reports from the data.

This paper will first describe the purpose and architecture of the HDI™ software, including the initial steps which have been taken in modifying the software so that it offers the functionality which will enable it to become a successful commercial product. The paper then describes some of the key functionality still to be developed.

## 2. Health Data Integration™

The Health Data Integration (HDI™) software [1] provides a general data integration framework with specific support for the needs of health data custodians and researchers. HDI™ takes a federated approach to integrating data sources, with web services being used to query heterogeneous data sources and retrieve the data for delivery to the user and further analysis.

HDI™ uses a meta-data layer to describe the data sources. This meta-data is used in building a user

interface which helps the user to explore the data. The meta-data is also used in the planning of complex queries across the data services which are then run by an executor service. Linking, reporting and analysis services provide extra functionality which health data custodians and researchers require to answer their questions.

Figure 1 shows a schematic of the HDI™ architecture, where data services can be distributed over the network and a set of HDI™ services are used to integrate the data for analysis by users in the Researchers Work Bench (RWB).

## 2.1 Privacy and Security

Privacy and security of the data which HDI™ integrates is paramount. A large part of building a product such as HDI™ is creating confidence amongst its target communities, particularly in respect to the legislative imperatives governing some of this data.

One such community is the large number of data custodians who maintain health related data sets. This is the case for clinicians who often maintain their own data sets relating to their patients. It is also the case for hospital administrators and Department of Health data custodians whose job is to record and maintain the data. It is vital that these data custodians have trust in HDI™ that their data will only be revealed to users who should have access to it.

Similarly, health policy makers and clinical researchers who will benefit most out of the linking of these data sources will need to feel confident that they are seeing all the data they can, without breaching any

confidentiality or security policies which apply to them.

**2.1.1 Authentication and Authorization.** HDI™ uses a token based approach to security of the data. Users are assigned different roles that then affect the

level of access they have to the data and other services. SAML [4] tokens are placed in the headers of the messages between services so that the role based access can be applied in each service and for each data source.

Currently the role based authorization is used only in gaining access to services or whole data sets. Future

work will involve more fine grained access to data and services and the use of more advanced policies to ensure security and privacy.

**2.1.2 Network Security.** All communications in HDI™ are done using SOAP messages. This is the case for messages between client and the services and service-to-services messages. HTTPS is used in the transfer of data between services, ensuring the highest level of security and privacy of the data.

**2.1.3 Encryption of identifying data.** To ensure privacy, HDI™ encrypts sensitive data before it is shipped for linking with other data. Using encryption to protect sensitive data for integration can be challenging: once encrypted, data can no longer be compared aside from exact matches. The data custodian decides on what data should be encrypted when they add a data source to HDI™.

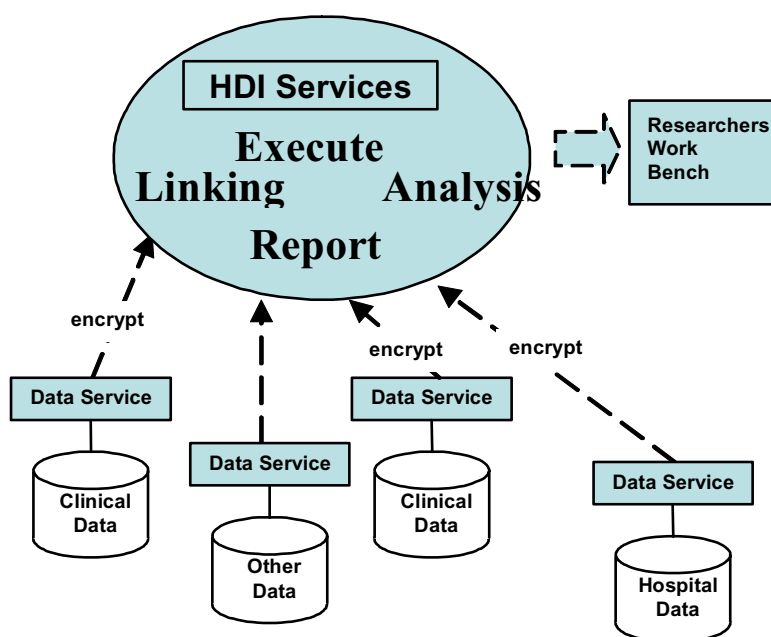


Figure 1. An overview of the HDI™ Architecture

## 2.2 Data Linking

Linking of patient records in different data sets, while maintaining the privacy of patients in the data sets, is core HDI™ functionality. This becomes more difficult when the identifying data is encrypted, and hence linking of encrypted data must be undertaken.

Further information about linking between data sets is contained in section 3 of this paper.

## 2.3 The Researchers Work Bench (RWB)

The Researchers Work Bench (RWB) is the current application based (as opposed to a web based) user interface for HDI™. The RWB provides the following functionality.

**2.3.1 Data Discovery.** The RWB offers the user a level of data discovery which allows them to investigate the data without the need to know in great detail the structure of the data sources.

Figure 2 shows the Query Developer, which allows users to select fields within the data sources they would like to query. Related information can then be selected from other data sources by drawing a link between fields in the data sources. Constraints can be added to any of the selected fields to restrict the records retrieved to a subset of records.

### 2.3.2 Analysis and Reports.

Analysis and reporting of the data must be easy for users to perform. A successful commercial version of

HDI™ will make it easy for users to go from linked data sets to standard analysis tools and reports.

HDI™ now contains standard medical analysis tools, such as the Kaplan Meier survival analysis and reporting functionality. In particular the reporting functionality is there to give novice users access to the data sets immediately, before they can start using the Query Developer to probe further into the data.

## 3. Data Linking

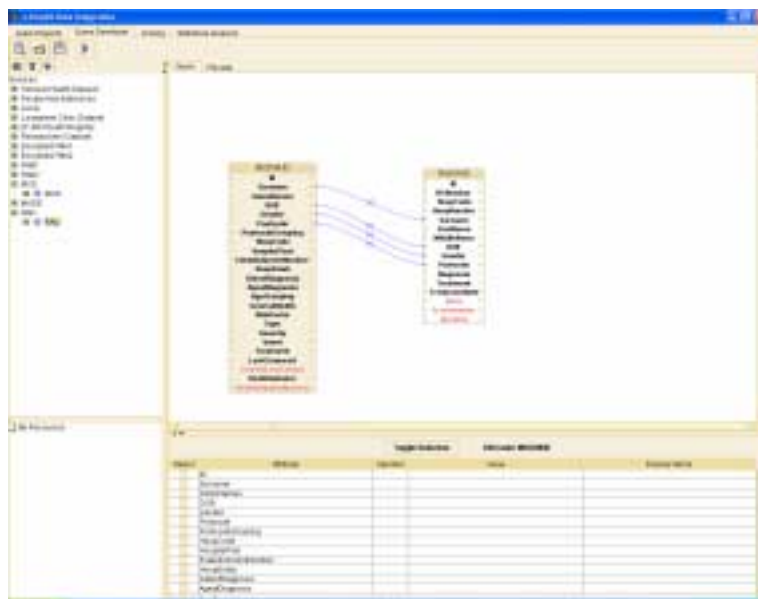
As mentioned above, linking of patient records while maintaining privacy provisions is core HDI™ functionality. The HDI™ team is researching different ways to improve the rates of matching once the data is encrypted. This research is currently investigating two different mechanisms, elaborative matching and use of distance functions.

The linking service is provided by HDI™ to retrieve data from different data services and then perform a linking algorithm on the data retrieved. The resulting links are stored in an internal data base and can be used as a link table between two data sets in the HDI™ user interface.

### 3.1 Elaborative Matching

In this algorithm extra fields are derived from existing data to correct for common data entry and conversion errors. For example, the input of a persons' date of birth is particularly prone to error and fields can be derived containing dates of birth with digits purposely switched.

These identifying fields are encrypted and sent to the linking service. The linkage server determines



**Figure 2. The Query Developer offers easy field selection and constraint specification**

which entries match based on a set of comparisons which is configurable by the HDI administrator.

### 3.2 Distance Functions

The distance function algorithm uses a feature-extraction encryption scheme which has two steps.

(1) Feature-extraction. At each data source a predefined feature function is determined by combining different distance functions for different types of data where their scores are above a certain threshold.

(2) Comparison: The encrypted features are then sent to the linking server which links records based on their distance scores above the thresholds.

The linkage server will use the comparison step to determine the set of matching records to be used when linking between data sets.

## 4. Integrating data sources

The addition of data sources to a data integration framework should allow for the seamless addition of the data into the data analysis functionality of the system. As well, the framework should support data standards which are specific to the target industry, thus making it easy for the user to know about the data without needing know about the data source.

### 4.1 Health Data Dictionaries

A number of clinical policy bodies are now starting to publish standard data dictionaries for data custodians to follow. While this helps standardize data sets and makes it easier to integrate these data, a number of data standards will exist across health areas. One such initiative is the National Cancer Control Initiative (NCCI) core clinical cancer data set [5], which can already be accessed in HDI™.

HDI™ allows for fields within data sources to be marked as being a field in the standard data set. This enables end users to know what the field contains, regardless of the name in the underlying data source.

The National E-Health Transition Authority (NEHTA) [6] is currently developing Clinical Data Standards and Terminologies. Once these data standards are published, they will be some of the data standards which are available within HDI™. New data sets which are added to a HDI™ network will then be able to publish their data with links between

their data set and the data standards. This will enable HDI™ to immediately run any of the standard Key Indicator Reports which have been developed for that standard data set against the new data sources.

### 4.2 Internal Meta-Data

The use of internal meta-data is a way for data custodians to add data sources once HDI™ is installed.

Currently the Data Custodian Module allows data custodians to add a data source by specifying the type of Relational Database Management System (RDBMS) and giving a link to the URL of the RDBMS. HDI™ then retrieves the schema information and allows the data custodian to specify information about each field and table in the schema.

The data custodian can specify to which field within a particular data dictionary the data links. This enables the data custodian to present a view of the data to end users which conforms to that particular standard data dictionary.

The data custodian also specifies how the data should be treated by HDI™, including

- whether the field should be published at all, and if not then this field or table will not be visible to the HDI™ user
- whether the field should be encrypted, in which case the field cannot be queried
- whether the field contains identifying data, in which case it is automatically used in the linking algorithm between two data sources

This internal meta-data is used within HDI™ to determine the level of access which users have to the data and how the data should be presented to the user.

## 5. The Commercialization Challenge

The requirements of developing software for research purposes to solve a particular problem are often very different to those of developing a software product. While there is great value in solving the problem, turning that software into a product which is a commercial success will always take more time and resources.

Most issues in turning research software into a commercial software package revolve around the robustness, scalability, usability and maintainability

of the software, along with the suitability of the software to meet the requirements of the users. As well as solving a central problem, the software must also provide functionality which allows the end user to easily access and use that core functionality. In the case of HDI™, the linkage of data sets while preserving privacy is the key piece of functionality, but the ease of querying, analyzing and reporting on the data may be the key to the success of a commercial version. With data integration products the quality of the administration tools for data custodians will also be of great importance.

The quality of the software, especially in dealing with health related data, is of the utmost importance. The HDI™ development team has established a Software Engineering Methodology and processes to ensure the quality of the software produced.

### 5.1 The importance of early adopters

The HDI™ software is of interest to Queensland Health to provide a way of linking the clinical data sets which have been collected by various clinicians and clinician groups to Hospital Admission and Radiology data. This will provide clinicians with invaluable information for clinical improvement as well as Queensland Health with information to drive health service provision and health policy.

Early work with Queensland Health revealed that HDI™ had the right core functionality with the linking of data sets, but had no mechanism to automatically report on the Key Indicators which the clinicians and Queensland Health were looking for.

Hence the first development task was to add a reporting module to HDI™ which would build some key reports about the data.

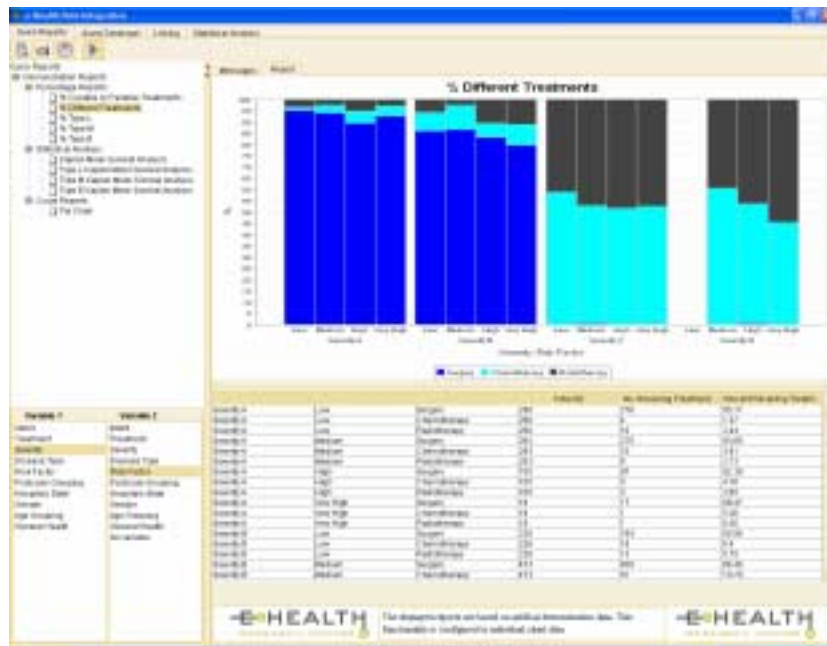
The EHRC has now installed an early version of HDI™ within Queensland Health for the Queensland Cancer Control and Analysis Team (QCCAT) to use with some available cancer data sets. The QCCAT and HDI™ teams

will continue to work together on the linking of cancer related data and development of reporting and analysis tools within HDI™.

Figure 3 shows the reporting tool which allows users to visualize pre-determined statistical functions run against their data.

### 5.2 Beyond early adopters

The further development of HDI™ depends not just on Queensland Health continuing their involvement, but on expanding the use of the technology into other organizations.



**Figure 3. A standard report on a test data set. This shows the percentage of patients who receive 3 different types of therapy when they have different disease severity and risk factor**

Besides the obvious commercial imperatives in gaining more customers of the HDI™ technology, HDI™ integration technology will improve greatly as the number of data sets it works with increases and the range of analysis and reporting it is required to do expands.

The EHRC has put in place an implementation team who are already working with other data custodians to integrate their data and provide analysis and reports.

**5.2.1 Number and size of data sets.** Adding more data sources to software products often leads to a non-linear increase in the complexity of the system, leading to a limit to the number of data sets which can be handled by the software. By using internal meta-data to describe how data interacts in the HDI™ system, adding data sets to HDI™ need not increase in complexity with each new data set.

**5.2.2 Other data sets.** Broadening the type of data sets which HDI™ has been used for to work with is imperative to ensuring that HDI™ will work with any data set that data custodians may have. Being able to link in environmental or genomic data sets would allow other factors contributing to health to be considered.

**5.2.3 Different Privacy Policies.** Privacy policies differ across jurisdictions and hence HDI™ will need to support a general privacy policy framework, such as XACML [8], to ensure that all policies are supported.

## 6. Conclusion

To improve our health care it is imperative that the heterogeneous data sources which are currently available are linked. This will give clinicians the ability to improve the treatment of their patients by

using the evidence for best practice currently locked in these data sets. Health policy researchers will be able to use the data to improve provision of health data and for policy research.

HDI™ is already proving a valuable tool for these groups of users. The challenge lies in building a complete solution which will be viable long term for a larger set of users.

## 7. References

- [1] K.L. Taylor, C.M. O’Keefe, J. Colton, R. Baxter, R. Sparks, U. Srinivasan, M. A. Cameron L. Lefort, “A Service Oriented Architecture for a Health Research Data Network”, *Proc. SSDBM ’04*
- [2] <http://www.e-hrc.net/>
- [3] L.G. Christine M O’Keefe, Ming Yung and R. Baxter. “Privacy-preserving linkage and data extraction protocol. In *Workshop on Privacy in Electronic Society in Conjunction with the 11<sup>th</sup> ACM CCS meeting, Washington DC, 2004*
- [4] Organization for the Advancement of Structured Information Standards (OASIS) “Security Assertions Markup Language (SAML) Version 1.1”, September 2003
- [5] <http://www.ncci.org.au/projects/data/dat01.htm>
- [6] <http://www.nehta.gov.au/>
- [7] Health Networks Build Medical Muscle, SOLVE, Issue 2, p8-9, Feb 2005.
- [8] Organization for the Advancement of Structured Information Standards (OASIS) “eXtensible Access Control Markup Language (XACML) Version 1.0”, February 2003