

MetaData Registry Implementation Project Experiences
The Wisconsin Department of Health and Family Services (DHFS)
Report to the MetaData Registry Coalition – 2000

INTRODUCTION AND OVERVIEW

The Wisconsin Department of Health and Family Services' (DHFS) mission is to "lead the nation in fostering healthy, self-reliant individuals and families." Toward this end, through its commitment to manage resources responsibly, DHFS has established a Data Stewardship Council. The members are directed to function as "stewards" of a common, shareable and valuable resource, as managers of a resource for the use and benefit of others. The MetaData Registry Implementation Project holds the promise of providing the Council with a "tool" that will help the stewards carry out their responsibilities. This progress report describes how MetaPro (beta version of the tool) has already begun to be incorporated in the work of the Council. At the end of the report, recommendations are noted that describe features the Department would find most helpful in the production version of MetaPro, scheduled for release in 2001.

ORGANIZATIONAL PERSPECTIVE

The Wisconsin Department of Health and Family Services is an "umbrella" organization that maintains over 450 manual and electronic information systems. These systems contain a wide array of identifying and other information about clients and the general citizenry of Wisconsin. As the principal health data source for the state, the DHFS has direct relationships to a number of federal agencies, such as the Health Care Financing Administration (HCFA) and the Environmental Protection Agency (EPA). Direct relationships exist between this Department and many other federal agencies as well. Additionally, the Department is a public sector member of HL7 and is actively involved in implementing relevant sections of the Health Insurance Portability and Accountability Act (HIPAA). The Information Technology and Resource Management Section (ITRMS) of the DHFS Bureau of Information Systems (BIS) is responsible for staff resources planning for HIPAA implementation, in support of the DHFS responsibilities for assuring compliance with HIPAA standards.

Information in storage ranges from vital permanent records to the most temporary of records, and from the most public of records to the most confidential. Although the State Registrar for vital records is an employee of the Department, responsibilities of that position are explicitly mandated under state statutes for the state record repositories containing birth, death, marriage, divorce and records of other vital events. Administrative records for social services programs such as Child Welfare, and benefits programs such as Medical Assistance, are maintained by the agency. The unit that administers the hospital discharge data is a part of this agency. Public health information related to the environment and occupational safety is maintained by this agency as well.

Although there is, and has been, nearly universal agreement among the program divisions and units within DHFS that name and birth date are primary to the identification of individuals, there is often surprising disagreement as to the name and birth date of record for the same person among the different systems. The DHFS Cross Program Participation Study (of persons participating in twenty programs during the period of July 1996 through June 1997) found that when a person was recorded in multiple systems, it sometimes was exceedingly difficult to locate all records related to that person. Often it was found to be difficult to locate records for an individual person within a single system. At times, different spellings of names and different dates of birth were recorded for the same person (in a single system).

In order to address these data consistency problems, in September 1998 a work group was formed to establish a core of data items for identifying individuals. Staff from across the department were invited to participate in the work group to develop a set of core (common/minimum) data elements that would best serve to *identify any individual* recorded in DHFS record keeping systems and/or served by DHFS programs. The work group was also directed to develop a supplemental data set to be used when core data elements alone would not identify an individual or when a database or file would not contain one or more of the core data elements. This work of the Department for identifying common core data elements for

achieving the unique identification of individuals was occurring at the same time as the national debate on a national system for a unique health identifier for individuals, as mentioned in the Health Insurance Portability and Accountability Act (HIPAA) of 1996.

STANDARDS DEVELOPMENT AND REGISTRATION

In September 1998, John Kiesow, Executive Assistant to the Secretary of DHFS, invited staff from across the department to participate in a work group. He charged the group with developing a set of core (common/minimum) data elements that would best serve to *identify any individual* recorded in DHFS record keeping systems and/or served by DHFS programs. The work group was also directed to develop a supplemental data set to be used when core data elements alone would not identify an individual or when a database or file would not contain one or more of the core data elements.

The Department's effort parallels the national effort to operate a metadata registry of industry-wide scope to support multiple interests, to load available metadata, to develop procedures for integrating metadata registry operations into department operations and to identify new requirements. Resource materials related to HIPAA, and available over the Internet, were utilized in the work group's effort at our state level.

A year later, in September 1999, the work group had completed its project and the DHFS adopted an initial set of Common Core Data Standards (Attachment A). Included in the introduction to the standards is the following paragraph:

*"High quality data integration, match/merge processing, and record linking depends heavily upon the existence of data necessary for unique identification of persons and organizations (e.g. clients and providers). At DHFS this data is classified into **Identifying and Corroborative Demographic Data**. Together, these data support many operational and analytical activities throughout DHFS. These are highly sharable resources and warrant standardization."*

In the second quarter of 2000, information from the March 2000 MetaData Coalition panel presentation at the DAMA International Symposium and MetaData Conference, entitled Implementing Meta Data Registries for Maximum Benefit, provided a thorough introduction to metadata registries. The panel consisted of Bob Mayes, Director, Information Systems Group, Health Care Financing Administration, David Patterson, Consultant, Data Union LLC, Kathleen Gundry, Technical Project Leader, SAIC and Judith Newton, Computer Specialist, NIST, the panel moderator.

Web links to the Environmental Protection Agency (EPA) and to the Australian Institute of Health and Welfare (AIHW) provided further access to conceptual information on metadata registries. The Power Point presentations on CD-ROM from the previous year's session provided information and direction well beyond the panel presentation. Information from the United States Health Information Knowledgebase (USHIK) internet site: <http://hmrha.hirs.osd.mil/registry>, also was quite helpful. This web site contained selected HIPAA data elements used by the X12 standards committee, the Health Level Seven (HL7) standards committee, the National Council of Prescription Drug Program (NCPDP), and the National Committee on Vital and Health Statistics (NCVHS). These sources were used in the analysis of the Department's Common Core Data Set, which originally was based upon hard copy materials from those committees. DHFS thus was able to build its approach on the concepts, standards, and practical approach to registering metadata that had been developed by the ISO/IEC JTC1 Working Group 2 o Subcommittee 32. The availability of the beta test version of the distributable MetaPro prototype metadata registry enabled DHFS to actively participate as a "beta tester" of the software and in the monthly meetings (via phone) of the Metadata Registry Implementers Coalition.

In June 2000, the Department's Data Stewardship Council was established. The initial charge of the Council was to guide the DHFS Bureau of Information Systems (BIS) in the continuing development and deployment of common core data standards. The Council was created to ensure that BIS actions in support of shared and well-managed data are based on program knowledge and priorities. The Council is to provide informed decision making in various areas of information management which are to include: standards setting, policy development, initiatives on data confidentiality and security, and shared application component specifications. Longer term, the Council will evolve and implement the concept of data stewardship. Those who control the Department's data are directed to do so as "stewards" of a common, shareable and valuable resource, as managers of a resource for the use and benefit of others.

In September 2000, the Information Technology and Resource Management Section (ITRMS) within the Bureau of Information Systems (BIS) began "registering" the initial set of data standards in the beta test version of distributable MetaPro. The ISO/IEC 11179 *Specification and Standardization of Data Elements* and the ANSI X3.285-1998 *Metamodel for the Management of Shareable Data* standards are being utilized as normative guides in this effort. The learning experience associated with the use of the MetaPro data registry builder to document data standards within DHFS, as well as the ability to obtain similar documentation on existing and evolving data element specifications at the national and other states levels, as well as data standards used abroad in other nations, is anticipated to be most helpful in describing the standards that have been adopted, as well as in refining the standards in the next version.

The Department's effort parallels the national effort of the USHIK to operate a metadata registry of the nation's health data with health industry-wide scope. This USHIK can serve to support multiple interests, to load available metadata, to develop procedures for integrating metadata registry operations into department operations, and to identify new requirements. From the state level, as represented through the Data Stewardship Council, the Department of Health and Family Services has direct relationships to the Health Care Financing Administration (HCFA) and the Environmental Protection Agency (EPA), development partners and sponsors of the distributable MetaPro. The metadata registry will support the data exchange relationships and agreements that exist between this Wisconsin Department and many federal agencies. Additionally, the Department is a public sector member of HL7 and is actively involved in implementing relevant sections of the Health Insurance Portability and Accountability Act (HIPAA). The ITRMS Section Chief has the responsibility for chairing the Stewardship Council, providing staff resources for ongoing activities and is responsible for staff resources planning for HIPAA implementation.

CORE DATA ELEMENTS

The work of the Department in building its data registry has begun with identifying common core data elements for the unique identification of individuals. Prior to participation in the (beta test) of MetaPro, data element entries in the 1999 United States Health Information Knowledgebase (USHIK) were studied. Of particular relevance was the data element "Personal/Unique Identifier - ID:0060-92-764-5622.00122400.v1". The definition section of the *Data Element Detail Report* for that element includes the statement; "Standards groups should be consulted regarding setting criteria for the recording of names." The comments section contains a brief discussion of the issues and mentions the California proposal. Reference is later made to the source document, *Core Health Data Elements - Report of the National Committee on Vital and Health Statistics*.

In the context of the discussion on the development of a national system for a unique health identifier for individuals, as included in the Health Insurance Portability and Accountability Act (HIPAA) of 1996, suggestions for alternative approaches were solicited by the National Committee on Vital and Health Statistics (NCVHS). Thirteen proposals were submitted by various interested parties for consideration for review by the NCVHS. These suggested approaches provided an overview of the strengths and weaknesses across a range of possibilities for the identifier. Five of the proposals were classified by the advisory committee as being alternatives to a unique patient identifier, and these proposals focused directly on core data elements necessary for unique identification. The appropriate resource materials for these proposed unique identifier approaches were available over the Internet, and were utilized in the ongoing work effort at our state level.

THE CALIFORNIA FHOP APPROACH

The California proposal submitted by the Family Health Outcomes Project (FHOP) from the University of California at San Francisco was examined in detail by DHFS. The FHOP study related to a virtual identifier or "black box" approach. The "black box" being a computerized probabilistic matching program that utilizes select demographic elements to form a matching key. The value of the key could be used as a common patient identifier.

Although the FHOP study was the basis for the proposal noted above, its original use was to develop linking a algorithm to construct a virtual identifier. This was in response to a California 1991 legislative mandate regarding Health Care for Women, Children, and Adolescents (AB99). The legislative committee recommended that the State of California adopt a set of data elements which would constitute a unique patient identifier for use by health and welfare programs. Standard Definitions of Data Elements as recommended by the California Interagency Collaboration Project is attached as Attachment A.

The proposal, developed by the Family Health Outcomes Project (FHOP) of the University of California, San Francisco, appeared to be most relevant to the question of what would constitute "a set of core (common/minimum) data elements that would best serve to *identify any individual* recorded in DHFS record keeping systems and/or served by DHFS programs."

Summary of the Family Health Outcomes (FHOP) Proposal

"The University of California, San Francisco Family Health Outcomes Project (FHOP) recommends the use of standard data sets for the identification of patient information. FHOP is part of the Department of Family and Community Medicine and is affiliated with the Institute of Health Policy Studies in California. FHOP has opted for data standardization and unique client identification instead of establishing a unique client ID. FHOP's identifying data elements consist of two sets namely Core Data Elements and Confirmatory Data Elements. The Core Data Elements consist of the following five (5) data items:

- 1. Birth Name*
- 2. Birth Date*
- 3. Birth Place*
- 4. Mother's First Name*
- 5. Gender*

The Confirmatory Data Elements consist of the following seven (7) data items:

- 1. Social Security Number*
- 2. Other Client Number*
- 3. Father's Name*
- 4. Mother's Maiden Name*
- 5. Current Name/Client Alias/Nickname*
- 6. County of Client's Residence*
- 7. Zip of Client's Residence*

The FHOP approach uses object oriented software technology and a method known as blocking technique. The blocking technique is used to determine the relative weighting of each of the common data elements and their sequence. From the resulting data set in their weighted order an alphanumeric string value is derived. This value is used to detect and link duplicate records in pilot projects which yielded impressive results. FHOP points out that the alphanumeric value based on the common core data elements can be used as a Common Patient Identifier. The Common Patient Identifier value can be destroyed after linkage. It will then serve as a Virtual Identifier. An object-oriented software matching algorithm is used for a probabilistic matching. The FHOP proposal is aimed at facilitating database linkage among the various centers of care with data standardization. They do not replace the institution specific identifiers that are currently used at the various branches of the statewide health services for managing the patient encounter and record keeping."

California's study utilized 602,269 records from 1992 data in California's Automated Birth Certificate Data Base. They found that using the five core data elements (birth name, birth date, birthplace, mother's first name and gender), as the record key, resulted in a 99.843% successful unduplication rate. FHOP's implicit belief was that the California Automated Birth Certificate Database contained only unique birth records.

The discriminating power of using all five core data elements was determined by dividing the number of unique records, when using those elements as the key, by the number of total records in the data set (assumed to be unique). Using full name alone, FHOP's study produced a result that was interpreted as having a .95388 probability of being a unique key.

THE WISCONSIN DHFS EXPERIMENT

California's identifying group of five core data elements was tested against Wisconsin data by constructing a matching string containing the core variables. It was found that using a matching string consisting of: (first name + middle name + last name + gender + mother's first name + place of birth + date of birth) as a key indicated that there were 630,325 unique records out of the total of 630,489.

California, in their study had found that using all five elements resulted in an unduplication probability of .99843 against the California database. The same elements against Wisconsin data resulted in a probability of being a unique key of .99974. Further, Wisconsin tests (on Wisconsin data) showed that using only two elements, a full birth name coupled with birth date, performed as well as five core data elements. That was surprising.

Findings between the two studies were nearly identical where they could be directly compared, and were complementary when answers to somewhat different queries were analyzed, as shown by the following table.

KEY Total Records Number of p(Unduplicate)*
Unique Records

Five Core Elements (Full Name, Birth Date, Gender, Place of Birth, Mother's First Name)

Wisconsin	630,489	630,325	0.99974
<i>California</i>	<i>602,269</i>	<i>574,492</i>	<i>0.99843</i>

Four Core Elements (Full Name, Birth Date, Gender, Place of Birth)

Wisconsin	630,489	630,325	0.99974
------------------	----------------	----------------	----------------

Three Core Elements (Full Name, Birth Date, Gender)

Wisconsin	630,489	630,325	0.99974
------------------	----------------	----------------	----------------

Two Core Elements (Full Name, Birth Date)

Wisconsin	630,489	630,325	0.99974
------------------	----------------	----------------	----------------

One Element (First + Middle+ Last)

Wisconsin	630,489	622,110	0.98671
------------------	----------------	----------------	----------------

One Element(First + Last Name)

<i>California</i>	<i>602,269</i>	<i>574,492</i>	<i>0.95388</i>
-------------------	----------------	----------------	----------------

Frequencies for Wisconsin Only

Last Name	630,489	93,857	0.14886
First Name	630,489	40,172	0.06372
Middle Name	630,489	34,340	0.05447
Date of Birth	630,489	3,287	0.00521

Ted Ohlswager, Section Chief (Special acknowledgement to Ted for his vision and data modeling expertise.)
Mike Nardi, Data Services Supervisor
Kathy Sterr-Macke, Data Administrator
Fred Buhr, Data Administrator-Consultant

One of DHFS' recommendations is to add a registry element (with memo fields) for ease of access to complete copies of reference and/or background documents. This recommendation results from a felt need, when using distributable MetaPro, to be able to quickly retrieve artifacts related to evolving national or local standards. Illustrative of this need, it would be most helpful to be able to quickly access and retrieve both the national and local artifacts relating to the DHFS classification scheme: *Identifying Data for Persons*. Since that scheme contains such a (deceptively) simple listing of data elements relating to first, middle, last name and birth date, source background documents are essential for understanding the context (and complexity) of the classification components.

CURRENT EXTENT OF REGISTERED DATA ELEMENTS IN BETA METAPRO

In testing the beta version of MetaPro, DHFS has identified itself as a local registry authority and has registered forty-seven data elements (as literally as possible) from the first set of Common Core Data Standards. In order to introduce minimal modification to the published standards, an electronic cut and paste approach was utilized for data entry.

A summary of the number of entries follows:

Classification Schemes	8
Conceptual Domains	12
Data Element Concepts	8
Data Elements	47
Value Domains	68
Permissible Values	
Added	75
Reused	4720

The reused codes primarily relate to Address standards and ISO and/or USPS codes and names for countries and states. New additions primarily relate the classification of identifying data for persons and service providers.

The top down approach to registering data elements was first tested. However, the bottom up "cut and paste" approach subsequently was utilized (and is the entry approach of choice for Wisconsin). Up to this point, object classes have not been explored and only a beginning approach to rules has begun. Illustrative of the bottom up approach are the following select entries for data elements and value domains related to the "Identifying Data for Persons" scheme included as Attachment B.

Using knowledge gained through the entry and discussion of the original set of standards into MetaPro, recommendations for the registration of additional data element standards will be presented to the Data Stewardship Committee at their February 2001 meeting. These standards are attached as Attachment C. Comparison of the original with the revised draft will provide an overview of the progress made in understanding the concepts and processes involved in the registry of standards. It should be noted that an ad hoc construct of "business concept" has been added

Since responsibilities for assuring compliance with HIPAA standards as well as governance aspects of metadata specific to DHFS are assigned to the Information Technology and Resource Management Section, a capacity to utilize the final production version of distributable MetaPro in a harmonizing role will be very valuable. What could be of particular value in accelerating the achievement of the administrative simplification intent of the HIPAA legislation would be to pre-load the forthcoming production version of distributable MetaPro with metadata for the data elements specified in the standards adopted for HIPAA.

DHFS appreciates the contributions of the MetaData Coalition in disseminating the concepts and practical experience with registering metadata, and alerting state agencies to the opportunity to actively participate as a "beta tester" of MetaPro, and to participate via phone connection in the information exchange of the nearly monthly meetings of the Coalition. It was information from the MetaData Coalition panel presentation at the DAMA International Symposium and MetaData Conference in March 2000 entitled **Implementing Meta Data Registries for Maximum Benefit** that provided the introduction to metadata registries. The panel consisted of:

Bob Mayes, Director, Information Systems Group, Health Care Financing Administration
David Patterson, Consultant, Data Union LLC
Kathleen Gundry, Technical Project Leader, SAIC and
Judith Newton, Computer Specialist, NIST (moderator).

Power Point presentations on CD-ROM from last years' session provided information and direction well beyond the panel presentation. Information from the United States Health Information Knowledgebase (USHIK) internet site: <http://hmrha.hirs.osd.mil/registry> which contained select HIPAA data elements used by the X12 standards committee, the Health Level Seven (HL7) standards committee, the National Council of Prescription Drug Program (NCPDP), and the National Committee on Vital and Health Statistics (NCVHS) was utilized in the analysis of the Department's Common Core Data Set (that was originally based upon hard copy materials from those committees). Links to the Environmental Protection Agency (EPA) and the Australian Institute of Health and Welfare (AIHW) provided further access to conceptual information on metadata registries and directly resulted in the pilot beta use of distributable MetaPro.

Since responsibilities for assuring compliance with HIPAA standards as well as governance aspects of metadata specific to DHFS are assigned to the Information Technology Resource Section, a capacity to utilize distributable MetaPro in a harmonizing role would be very valuable. Pre-loading the production distributable MetaPro with HIPAA standards and other standards would be of immense value.

RECOMMENDATIONS FOR THE PRODUCTION VERSION OF DISTRIBUTABLE METAPRO
DHFS looks forward to the release of the production version of MetaPro in 2001 and ongoing involvement with the Coalition in the development and documentation of metadata. Following are recommendations concerning the production version of MetaPro:

1. Pre-load all United States Health Information Knowledge Base (USHIK) metadata as well as all metadata included in EPA's hosted version of MetaPro. **, a capacity to utilize distributable MetaPro in a harmonizing role would be very valuable. Pre-loading the production distributable MetaPro with HIPAA standards and other standards would be of immense value.**
2. Develop an ad hoc reporting search and retrieval capacity that would provide the ability to select data elements and generate a comparison matrix that could readily be understood by knowledgeable administrative staff, such as those participating on a Data Stewardship Council.
3. **Add a metadata element that would describe the degree of "confidentiality" attached to components of registered data collections and databases.** Wisconsin's findings concerning the unduplicating power of complete name and birth date, along with its observations concerning time sensitive linkage, may suggest a discussion by the Coalition of an object oriented metadata datatype: one having a name and birth date public interface along with a blood sample (blob) private interface. Such a datatype may be a candidate for the unique patient identifier referenced earlier in relation to the United States Health Information Knowledgebase (USHIK).
4. **Add a registry element (with memo fields) for storage and retrieval of complete copies of reference and/or background documents.**

5. Develop a search function that would permit “drill down” comparison of standards (i.e. HIPAA related) to the permissible values and format levels.
6. Develop an online help system that would incorporate a tutorial for developing metadata and registering legacy systems and data collections.