

# Pay-as-You-Go Information Integration

David Maier, Nicolas B. Rayner  
Department of Computer Science  
Portland State University

## ABSTRACT

We propose to incrementally elaborate a dataspace through the stages of characterization, customization and checking. After describing each of these stages, we present a specific example from the domain of medication standards.

## 1. INTRODUCTION

Recently, *dataspace management* has been proposed as an approach that takes a holistic view of all information in an enterprise, be it structured, semi-structured or unstructured, and whether or not it is supported by an explicit information system [1, 2]. While its goal is to provide services over the entirety of the data, dataspace management takes the pragmatic view that initial limits on time and effort may only permit simple services at first, such as cataloging and keyword search. Additional services or capabilities can come later, little by little, as resources permit. Such a pay-as-you-go approach seeks a steady return on investment (ROI), rather than a long period of implementation before any services become available.

## 2. CHARACTERIZATION AND CUSTOMIZATION

We have been investigating one path to pay-as-you-go elaboration of a dataspace, involving tools to aid in incremental characterization and customization of a dataspace. We assume that the data sources in a dataspace may be incompletely – or incorrectly – documented.

The first stage is to run routines that determine a class of characteristics or traits of one or more data sources. These can be fairly simple, as in current data profiling systems, such as compiling value distributions for fields and determining keys, or more complex, such as detecting domain-specific structure in generic representations.

The second stage presents customizations or enhancements that are enabled by specific discovered characteristics. Consider, for example, a UMLS-style generic relationship structure with **related(Concept1 Rel\_Name Concept2)** as schema. Suppose characterization discovers that in every tuple  $\langle c1, \text{has-form}, c2 \rangle$ , concept  $c1$  is always a *clinical drug* and  $c2$  is always a *dose form* (e.g., tablet, capsule, syrup). Then one possible customization is to factor out tuples of this form from **related** into a specialized table **hasForm(ClinicalDrug, DoseForm)**, either in materialized or virtual form.

The third stage is to generate a check condition that can monitor that the relevant characteristic still holds as data sources evolve, and hence that the customization is still valid. For example, if a new version of the **related** source contains an additional tuple  $\langle c1, \text{has-form}, c2 \rangle$ , where  $c1$  is a *branded drug*, then we must reconsider the **hasForm** factoring.

We imagine a framework for dataspace elaboration where new modules can be plugged in, each with characterization, customization and checking components. Call such an extension a “C-C-C” module.

## 3. A MEDICATION DATASPACE

We give a more detailed example from a dataspace we are currently exploring. The RxSafe project (led by Samaritan North Lincoln Hospital and Oregon Health & Science University) is developing a consolidated medication-list facility for rural elders in Lincoln County, Oregon. We are investigating various standards proposed for e-prescribing to add value to RxSafe, such as noting equivalences of generic and brand names, or grouping medications by drug class. Two particular sources in our dataspace are RxNorm and NDF-RT. RxNorm [3] is an effort from NLM to produce a standard nomenclature for medications and their components. NDF-RT (National Drug File – Reference Terminology) [4] contains drug-class information (among other things) developed by the US Department of Veterans Affairs.

We would like to connect information from these two sources to link brand names with drug classes. Figure 1 shows an excerpt from a table we derived from RxNorm, relating Semantic Clinical Drug (SCD) with brand name. (SCD is essentially the most complete generic name for a drug, giving ingredients, their strengths and a dose form.) There are 8,431 SCDs in this table. Figure 2 is an excerpt from a table derived from NDF-RT relating SCDs with drug class and class type. (Class types are more general categories over drug classes.) This table has 6,661 entries. These two tables are themselves the results of (manual) characterization and customization of the RxNorm and NDF-RT data sources.

It would seem that a join of these two tables would connect brand names and drug classes for us. The problem is that the connection is incomplete. About 54% of the SCDs in the RxNorm table do not appear in the NDF-RT table. Going in the other direction, 42% of the SCDs in the NDF-RT

table are missing from the RxNorm table. (Most of them are in RxNorm, just not connected with any brand name.) However, the situation is probably not as severe as it sounds. Figure 3 illustrates what we think is happening – there are variations in strength and dose form.

clinical_drug	brand_name
Acarbose 50 MG Oral Tablet	Precose
Acebutolol 400 MG Oral Capsule	Sectral

**Figure 1:** SCD-to-brand-name connection derived from RxNorm.

rxnorm_scd	ndfrt_class	class_type
Acarbose 50 MG Oral Tablet	HYPOGLYCEMIC AGENTS, ORAL	HS
Acebutolol 200 MG Oral Capsule	BETA-BLOCKERS/RELATED	CV
Acebutolol 400 MG Oral Capsule	BETA-BLOCKERS/RELATED	CV

**Figure 2:** SCD-to-drug-class connection derived from NDF-RT.

A human could probably figure out the connection between brand names and drug classes on a case-by-case basis. But is there possibly a C-C-C module that might overcome this connection problem? We think so. We could start with the collection of all SCDs in RxNorm that do have drug-class information in NDF-RT. We could then run a characterization routine to see if this collection obeys any non-trivial *stratification*: an equivalence relation on a domain where equivalent items have the same image under a relationship. In our example, there may be a stratification based on equality of ingredient lists, ignoring strength and dose form. (Note that we might require a prior customization to pick apart an SCD in order to express this equivalence.)

If there are such stratifications, we could choose one to help connect “unconnected” brand names with drug class. That is, consider a brand name b1 that is connected in RxNorm to an SCD s1, but s1 is not assigned a drug class in NDF-RT. If s1 is equivalent to another SCD s2 that does have a drug class c2, then we can impute c2 as the class for

b1. (Note by the nature of a stratification, any SCD equivalent to s1 will yield the same class c2.) The check condition for this customization is that the particular stratification continues to hold.

Classified in NDF-RT	Linked to Brand Name in RxNorm
	Acebutolol 100 MG Oral Capsule
Acebutolol 200 MG Oral Capsule	
Acebutolol 400 MG Oral Capsule	Acebutolol 400 MG Oral Capsule
	Acebutolol 400 MG Oral Tablet
	Acebutolol 5 MG/ML Injectable Solution

**Figure 3:** Example of missing connections between Figures 1 and 2.

#### 4. CURRENT WORK

We are currently developing specific C-C-C modules, as well as considering high-level ways to define them and a framework for incrementally investigating and enhancing a dataspaces.

#### 5. ACKNOWLEDGMENTS

This work is supported by NSF grant IIS-0534762 and AHRQ contract AHRQ 1 UC1 HS014928-01.

#### 6. REFERENCES

- [1] A. Y. Halevy, M. J. Franklin, D. Maier. Principles of dataspaces systems. *Proc. of the Twenty-Fifth ACM SIGACT-SIGMOD-SIGART Symp. on Principles of Database Systems*, Chicago, June 2006.
- [2] M. J. Franklin, A. Y. Halevy, D. Maier. From databases to dataspaces: A new abstraction for information management. *SIGMOD Record* 34(4), December 2005.
- [3] S. Liu, W. Ma, R. Moore, V. Ganesan, S. Nelson. RxNorm: Prescription for electronic drug information exchange. *IEEE IT Professional* 7(5), September 2005.
- [4] S. H. Brown, et al. VA National Drug File Reference Terminology: A cross-institutional content-coverage study. *Proceedings from the Medinfo 2004 World Congress on Medical Informatics*, San Francisco, August 2004.