Integrating Existing Drug Formulation Terminologies Into an HL7 Standard Classification using *Open*GALEN

Chris J. Wroe, M.B. BChir[1], James J. Cimino, M.D.[2], Alan L. Rector M.D. PhD[1]
[1] Medical Informatics Group,
Department of Computer Science, University of Manchester, UK
& OpenGALEN
[2]Columbia University Department of Medical Informatics, New York, New York

Many terminologies exist for the form of drugs - i.e. tablets, capsules, sprays, suppositories, etc. However, they have surprisingly different content. To communicate medication messages effectively, a mechanism is needed to translate between these existing terminologies. An ontological approach, based on techniques developed by OpenGALEN, has been used to build a drug form terminology for HL7 version 3. It integrates existing terminologies from commercial drug information vendors and regulatory authorities, and provides a framework for translating between them. To date, term sets have been included from First DataBank, the FDA, Multum and Micromedex, to produce a terminology of 820 concepts. The approach is made practical by distributing the knowledge engineering effort to volunteers with experience of the domain, and then integrating the knowledge into a logically consistent classification.

INTRODUCTION

Drug formulations – *i.e.* the specification of the form, intended route, and method of administration – is a seemingly small and well-defined domain. Developing a consistent standard terminology for such a domain would seem a simple matter. However, the experience of HL7 in trying to produce a common terminology based on a variety of sources from drug information vendors and authorities showed it to be surprisingly difficult. This paper describes how the use of *Open*GALEN's techniques which combine a simple, easy to use 'Intermediate Representation' with an underlying description logic (GRAIL) allowed a uniform approach to this problem. It illustrates that even seemingly simple domains can be surprisingly complex, and the value of a formal approach with automatic classification in dealing with those complexities relatively simply and quickly.

Health Level 7 (HL7)[1] version two is the *de facto* standard for many forms of healthcare messaging. Version 3 is currently under development with the aim of providing more systematic support for standardised message transfer between health systems based on underlying standard models and vocabularies. HL7 messages specify the accepted term set to be used in each attribute of the model and hence each field of the message. Large domains such as 'clinical drug', 'anatomy', or 'condition' use pre-existing externally maintained vocabularies. However, for smaller domains, such as drug formulations, HL7 develops and maintains its own term sets.

Many systems require access to the formulation component of a medication message, for example to find equivalent products to the one specified or to trigger formulation specific decision support rules. Existing drug knowledge base vendors have created their own formulation vocabularies. One of the co-authors Cimino found the number of exact matches between term sets strikingly low[2]. HL7 therefore began developing a unified drug formulation classification to provide a common point of reference for medication messages. Even though the size of the unified vocabulary was kept small (about 150 terms) development of the hierarchies proved more difficult than expected.

- Drug form terms in common usage ('oral tablet', 'metered dose inhaler') are actually complex pre-coordinated concepts which include additional information about intended route of administration and associated actions. This information can be represented in a classification by allowing a concept to have more than one parent, e.g. 'oral tablet' would be a child of both 'oral form' and 'tablet form'. Ensuring all parent child relationships are present proved difficult even for a classification with 150 concepts.
- The unified classification must be related in some way to existing formulation vocabularies. There are numerous existing vocabularies, which cover a wide range of granularities, e.g. '*aerosol'* to '*pressurised aerosol suspension, breath activated'*

*Open*Galen

*Open*GALEN[3] is a not-for-profit organization, grown out of the GALEN consortium to make available the techniques and resources developed during GALEN[4] and GALEN-In-USE[5] projects.

*Open*GALEN's ontological approach to terminology management is to define *what* concepts mean, rather than *where* they should appear in hierarchies. These formal concept definitions are then available to software applications to interpret messages expressed using these concepts. For applications that require a hierarchical organisation of concepts, classification software can automatically produce hierarchies customised to the specific application's requirements. We still must create hierarchies of the elemental concepts used to compose the definitions. However these hierarchies *are* maintainable by hand because of their smaller size and simpler mono-axial structure[6].

To take an example, the pre-coordinated form concept 'oral tablet' consists of the elemental concepts 'tablet' and 'oral'. These concepts are manually arranged in independent mono-axial form and route hierarchies. The classification software (based on description logic) then uses the concept definition together with these simple hierarchies to infer where 'oral tablet' should be placed in a multi-axial hierarchy. By selecting the abstract grouping concepts, such as 'oral form', the hierarchical structure can be customised to support a specific task, whilst still remaining consistent with a common reference source.

Intermediate Representation and GRAIL

Most concept definitions in *Open*GALEN are authored in *Intermediate Representations* which is easy to learn and tolerant of minor variations of style[7]. Each concept definition is called a *dissection* and consists of links between *descriptors* denoting more elemental concepts. Intermediate Representations are 'soft' and can be quickly tailored to specific requirements or individual preferences. Translation rules then specify how the definitions in the Intermediate Representations are expanded and normalised to standard constructs in the more formal GRAIL description logic[8].

The aims of using *Open*GALEN classification techniques in this project were:

- Expand the initial HL7 classification to incorporate existing commercial vocabularies.
- Distribute the knowledge engineering task to volunteers with experience in using the existing vocabularies.
- Allow the knowledge engineers to work directly in defining each term rather than indirectly encoding the knowledge in hierarchical structures.
- Provide a means of integrating the knowledge authored by the volunteers into one logically consistent classification.

METHODS

Two key knowledge resources are needed to support the *Open*GALEN approach.

1. Semantic definitions for every drug form concept. These can either be written directly in Grail, or more commonly an Intermediate Representation dissection, which is then expanded into Grail.
2. A collection of simple mono-axial hierarchies containing the sets of concepts used to compose the semantic definitions, e.g. oral, tablet, inhaler etc.

The development process centred on the knowledge engineering required to produce these two resources. A key aim is to distribute the knowledge engineering task to those with the most knowledge of the existing vocabularies. However, a degree of central knowledge engineering is needed by people with previous experience of the technique to start the process. The initial intermediate representation was that used in the Drug Ontology project[9]. The Drug Ontology is an ongoing UK NHS funded project to develop a knowledge base to support the PRODIGY project - a computerised drug prescribing support system[10].

Dissections were created centrally for concepts in the unified HL7 form vocabulary, together with a collection of mono-axial hierarchies for each category of concept used to compose the dissections. The classification software was then used to produce a multi-axial classification based on those dissections. The classification was distributed to volunteers together with the underlying concept definitions. This information acted as a set of examples from which volunteers could construct new definitions as needed.

Term sets were obtained from the FDA, First Databank, Multum and Micromedex. The First Databank set was divided in half, resulting in 5 term sets containing between 115 and 148 terms. Nine volunteers were recruited and each term set was distributed to four reviewers. The term sets included the name and code for each drug form, along with any definitions (FDA only). Reviewers were then asked to match the terms in their sets against those in the unified HL7 set and to classify their findings into one of four categories:

1. The term exactly matches a concept in the unified HL7 classification based on its concept definition.
2. A new and *complete definition* could be constructed using sections of example definitions.
3. Only a *partial definition* could be constructed.
4. No definition could be constructed.

The work from the volunteers was submitted to *Open*GALEN for integration. Terms exactly matching the HL7 concepts were assigned the same definition. Complete definitions authored by the volunteers were checked for legal syntax and then incorporated. Partial and missing definitions were completed usually by extending the semantic model used to represent the concepts and/ or expanding the elemental concept hierarchies. For example 'film coated tablet' required the extension of the representation to include the link 'HAS_PART' and the concept 'film coating'.

The complete set of definitions was classified automatically to produce a multi-axial hierarchy.

RESULTS

The unified term set developed by HL7 consisted of 164 terms and each was assigned an intermediate representation dissection. Preliminary mono-axial hierarchies were created for basic form, route, associated device, associated administration action and drug absorption concepts.

Nine reviewers returned a total of 16 evaluations. Some reviewers identified only the exact matches (9 sets) while others provided exact matches plus Grail descriptions for complete, partial and non-matches (7 sets). A comparison was made across the sets to determine inter-rater agreement on exact matches. In general, the terms identified as exact matches by one reviewer were either exactly the same as, or totally subsumed by, the set of terms identified as exact matches by other reviewers. When a discrepancy occurred, the majority rule was followed. In only one case did reviewers disagree about *what* the exact match was. This was due to a redundancy in the Galen set (elixer and elixir). As a result, of the 643 terms, 216 were classified as exact matches to a total of 136 *Open*Galen forms (71 matched one term, 50 matched two terms, 12 matched three terms, and 3 matched terms in all four source terminologies).

OpenGALEN 3000215 Effervescent tablets
fda|503|"TABLET, EFFERVESCENT"
fdb|9|"tablet, effervescent, oral"
mdx|70|effervescent tablet
multum|102|tablet \ oral \ effervescent

Figure 1. Consolidated results for the exact matches of *Open*GALEN effervescent tablets.

Figure 1 lists the exact matches for 'Effervescent tablets' in four existing term sets. It was agreed for example that the concept exactly matched the term 'TABLET, EFFERVESCENT' with a code of 503 from the FDA term set. Of note, there were five cases in which two terms from the same terminology set mapped to the same Galen term (e.g., "Mouthwash" and "Gargle" from one term set were both mapped to "Mouthwash/Rinse").

For 356 of the remaining 427 terms, new *Open*GALEN definitions were authored and classified as either a complete or partial match by the volunteers. Compared to the exact matches, more disagreement occurred between reviewers over the content of the new definitions (53 differences). There were two major causes for disagreement

1. A volunteer may find a very closely related concept in the example definitions and use that rather than looking further for an exact match or flagging it as only a partial definition. There was no equivalent concept present in the initial list for the term '*tablet, sustained action, oral*' from the FDB term set. Two volunteers independently constructed new concept definitions using the initial set as examples. Volunteer A chose 'sustained drug release' as the drug absorption term, while volunteer B chose 'slow drug release'. The completed definitions are shown in Table 1.

tablet, sustained action, oral		
Slot	Volunteer A	Volunteer B
	New description, complete match	New description, complete match
Basic Form	tablet	tablet
Route	oral	oral
Action	ingestion	ingestion
Drug Absorption	sustained drug release	***slow*** drug release

Table 1. Tabular summary of a new concept definition authored by two independent volunteers for '*tablet, sustained action, oral*'.

2. If only a partial definition could be constructed from the examples, one volunteer may categorise it as a partial match while the other extend the representation to define the concept and categorise it as 'complete with additions'. As for the example above, two volunteers independently constructed new concept definition for the term '*capsule sustained release, 24hr, hard, oral*' from the FDB term set. However, the concept 'hard' was not present in example definitions so only a partial definition could be completed. Volunteer A constructed this definition and flagged it as a partial match. Volunteer B added the concept 'hard' and used it to construct a complete definition.

capsule sustained release, 24hr, hard, oral		
Slot	Volunteer A	Volunteer B
	New description is partial match	New description is complete match using additional terms
Basic Form	capsule	capsule
Route	oral	oral
Action	ingestion	ingestion
Drug Absorption	24 hour extended release	24 hour extended release
Feature	--	***hard***

Table 2. Tabular summary of a new concept definition authored by two independent volunteers for '*capsule sustained release, 24hr, hard, oral*'.

The schema for the intermediate representation was extended to allow definition of all drug form concepts categorised by the volunteers.

In total 835 concept definitions were derived from the volunteers' work and the unified HL7 term set. These were specified as intermediate representation dissections to allow easier analysis and maintenance. Table 3 shows the average frequency of each semantic link used within dissections created for each source vocabulary. Of note, only a minority of FDA concepts contain route information, in contrast, to FDB concepts. FDB terms also contained more semantic information as a whole within the formulation concept. Micromedex and Multum had intermediate levels of semantic information.

Semantic links	Average number of links per dissection			
	FDA	Medex	Multum	FDB
Basic formulation	1.00	1.00	1.00	1.00
Route	0.11	0.30	0.91	0.91
Associated action	0.29	0.40	0.46	0.52
Pre-delivery Formulation	0.10	0.12	0.10	0.20
Associated device	0.04	0.11	0.11	0.20
Drug Absorption	0.17	0.13	0.04	0.09
Miscellaneous feature	0.03	0.00	0.09	0.15
Other semantic links	0.28	0.28	0.19	0.31
Total	2.02	2.34	2.90	3.38

Table 3. Average number of links per dissection for concept definitions from each existing vocabulary.

Manually created mono-axial hierarchies

Each concept used within a dissection was organised into one of a collection of simpler hierarchies. Table 4 shows the size and examples for each hierarchy used. For example there were 73 basic drug formulations such as 'Capsule basic drug formulation', 36 different routes such as 'oral', etc

Elemental hierarchy	Hierarchy Size	Example term
Basic drug formulation	73	Capsule basic drug form
Route	36	Oral
Actions	25	Injecting
Component substances	22	Coated particles
Associated devices	15	Inhaler device
Drug absorption	14	Slow release
Formulation functions	12	Effervescing
Misc. features	6	Soft

Table 4. Elemental hierarchies used in definitions

Delivery

The 820 concept descriptions were expanded from intermediate representation to GRAIL and automatically classified into multi-axial hierarchies. To allow easier browsing, the concepts were grouped into subsets and smaller hierarchies produced based on specific routes or basic formulations, e.g. a *tablet* hierarchy and an *oral form* hierarchy. These were then made available as a set of web pages.

Every concept in the hierarchies was linked to its definition. It was also colour coded to indicate the source of each term. A concept in the hierarchy could actually represent two or more equivalent concepts originating from different vocabularies. A number in angle brackets denoted the number of equivalent terms. For example, the '<3>' following 'Delayed Release Tablet' in Figure 2 shows it is one of three equivalent concepts. The (OG) in brackets denotes the displayed term is from the *Open*GALEN term set. Following the link to the definitions would show that it is equivalent to 'delayed-release tablet' from Micromedex and 'tablet, delayed release, oral' from First Databank.

```
Tablet - route unspecified' <3> (OG)
 .  controlled-release tablet' (Medex)
 . .  TABLET, DELAYED RELEASE' (FDA)
 . . .  Delayed Release Tablet' <3> (OG)
 . .  TABLET, EXTENDED RELEASE' <2> (FDA)
 . . .  TABLET, FILM COATED, EXTENDED
                                RELEASE' (FDA)
```

Figure 2. A small extract of the hierarchy for *tablet* drug forms. The complete hierarchies are available at http://www.cs.man.ac.uk/mig/people/wroec/hl7material.htm

DISCUSSION

The ontological approach of using knowledge about concepts to organise and manage terminologies is not unique to this project. Similar techniques have been used in the development of large terminologies such as SNOMED RT[11], SNOMED CT[12] and MED[13]. Environments have also previously been developed to support distributed development of these kinds of terminologies[14].

This project demonstrates that even for much smaller vocabularies in the order of hundreds of concepts, a formal ontological approach is practical and necessary for consistent results. This project also introduces a viable methodology for integrating *existing* commercial terminologies.

Terminologies can no longer be seen only as supporting the internal representation of information in isolated systems. The aggregation of smaller healthcare organisations has forced disparate health IT systems to communicate.

We have developed a methodology using an ontological approach, which allows the integration of existing terminologies within disparate systems. This was achieved by distributing the knowledge engineering task to those with most experience of using the existing vocabularies and then integrating the knowledge into one cohesive resource.

The results show that the ontology-based approach provides a practical framework for capturing the human knowledge needed to map between one terminology and another – the inter-rater agreement was high for exact matches.

However, even for small terminologies, the degree of variability is striking. Across the drug form terminologies: over half of the terms had no exact counterpart in other terminologies. This is due, in part, to differences in the degree to which associated attributes such as intended route are included and also to the level of detail in some terminologies ("pressurized aerosol suspension, breath activated" vs. "Aerosol"). In these cases, the generated hierarchies provide a mechanism to find the nearest mapping. By comparing the semantic definitions for the two concepts a measure of the degree of disparity between the two concepts can also be found.

In addition to allowing translation *between* terminologies, the method also provided a framework for identifying redundancy *within* individual terminologies. The method helped identify internal redundancy in 5 instances - and these were from reviewers who worked for the companies in question.

No specific training or specialised tools were provided to the volunteers. All the volunteers' work was guided by the example set of concept definitions, and direction of one of the co-authors. Despite this, a satisfactory concept definition was captured for the majority of terms. The low levels of inter-operator agreement for complete and partial matches did however demonstrate that this minimalist solution to training and support was not completely sufficient. It is hoped that in future projects we can develop a simple set of tools based on those used within *Open*GALEN, which will provide more support for remote knowledge engineers.

ACKNOWLEDGMENTS

With thanks to the volunteers: Steve Brown, Julia Kim, Sandra Boyer, Douglas Martin, Pavla Frazier, Tom McDonald, Terri Merridith, and Shaun Shakib who performed the knowledge engineering for the existing drug form terminologies. Also thanks to Stan Huff and Ed Hammond for inspiring the work.

The initial conceptual schema was that developed in collaboration with the Sowerby Centre for Health Informatics, during the Drug Ontology project funded by the UK NHS Executive. The GALEN Consortium contributed to the foundations of the *Open*GALEN methodology during the GALEN and GALEN-IN-USE projects funded by the European Commission.

REFERENCES

1. Health Level 7, HL7 Home Page, www.hl7.org
2. Cimino J.J., McNamara T.J., Meredith T., Broverman C.A., Eckert K.C., Moore M., Tyree D.J. Evaluation of a Proposed Method for Representing Drug Terminology. Proc AMIA Symp 1999;:47-51.
3. *Open*GALEN, www.openGalen.org
4. Rector A, Nowlan W, Glowinski A. Goals for Concept Representation in the GALEN project. In: 17th Annual Symposium on Computer Applications in Medical Care (SCAMC-93); 1993: McGraw Hill; 1993. p. 414-418.
5. Rector A., Zanstra P.E, Solomon W.D., Rogers J.E. Reconciling Users' Needs and Formal Requirements: Issues in developing a Re-Usable Ontology for Medicine. IEEE Transactions on Information Technology in BioMedicine, Special issue on the EU Healthcare telematics programme; Vol. 2 No.4 pp. 229-242
6. Rector A. Coordinating taxonomies: Key to re-usable concept representations. In: Barahona P, Stefanelli M, Wyatt J, eds. Fifth conference on Artificial Intelligence in Medicine Europe (AIME '95). Pavia, Italy: Springer, 1995:17-28.
7. Solomon, W., A. Roberts, J. E. Rogers, Wroe C.J., A. L. Rector. Having our cake and eating it too: How the GALEN Intermediate Representation reconciles internal complexity with users' requirements for appropriateness and simplicity. Proc AMIA Symp 2000;: 819-823.
8. Rector A, Bechhofer S, Goble C, Horrocks I, Nowlan W, Solomon W. The GRAIL concept modelling language for medical terminology. Artificial Intelligence in Medicine 1997;9:139-171.
9. Solomon D.S., Wroe C.J., Rogers J.E., Rector A. A reference terminology for drugs. Proc AMIA Symp 1999;:152-155
10. Purves, I.N. PRODIGY: implementing clinical guidance using computers. British Journal of General Practice (1998);48:1552-1553
11. Spackman, K.A., K.E. Campbell, and R.A. Côté, SNOMED-RT: A reference Terminology for Health Care. Proc AMIA Symp 1997;: 640-644.
12. SNOMED CT, http://www.snomed.org/snomedct_txt.html
13. Cimino JJ, Clayton PD, Hripcsak G, Johnson SB. Knowledge-based approaches to the maintenance of a large controlled medical terminology. JAMIA. 1994; 1(1): 35-50
14. Campbell KE, Cohn SP, Chute CG, Shortliffe EH, Rennels G. Scalable methodologies for distributed development of logic-based convergent medical terminology. Methods Inf Med 1998; 37(4-5):426-39