ICBO: International Conference on Biomedical Ontology
July 28-30, 2011 · Buffalo, NY, USA

Biomedical Analyses: OWL Model Based Edition

Pierre-Yves Vandenbussche[1,2], Ferdinand Dhombres[1], Sylvie Cormont[3], Jean Charlet[1,3], Eric Lepage[3]

[1]INSERM UMRS 872 ÉQ.20, Paris, France
[2]Mondeca, Paris, France
[3]AP-HP Assistance Publique – Hôpitaux de Paris, Paris, France

Abstract. Background and Objectives. The Assistance Publique Hôpitaux de Paris (Public hospital of Paris and its suburbs; APHP) developed a biology dictionary independent from laboratory management systems (LMS). This dictionary is interfaced with the international nomenclature Logical Observation Identifiers Names and Codes (LOINC), and developed in collaboration with experts from all biological disciplines. We aim to establish a platform for publishing and maintaining the APHP laboratory data dictionary, which can satisfy both the requirements concerning the controlled vocabulary and those related to maintenance processes and distribution. **Material and Methods.** Data complexity and data volume show the need to establish a platform dedicated to the terminology management. This replaces the use of a spreadsheet tool that might show weaknesses. After describing the dictionary, we identify requirements for the nomenclature management, and the inadequacy of existing software. Our method is based on the design of a OWL hub meta-model supervising organization systems. **Results.** We describe how the modeling, data migration and integration/verification steps in the new tool were used to meet these requirements. The core of our work is based on the modeling effort which integrates multiple dimensions: (i) interoperability regarding data exchange standards, and (ii) dictionary evolution. This model has been implemented in the APHP context. Structuring data representation has led to a significant data quality improvement.

1 Introduction

One of the projects of the Assistance Publique-Hôpitaux de Paris [1] (AP-HP) new information system is to acquire a biological analysis dictionary (AnaBio) common to the whole production chain: prescription, analysis processing in the laboratory management systems (LMS) and transmission of the result. Useable by all the LMS, active in the 45 hospitals of the institution spread into 165 laboratories, the repository on which is based this dictionary should ideally remain independent of these tool constraints. The dictionary achieved offers the managerial flexibility necessary to its daily use while maintaining the semantic interoperability with other international health organisms through its alignment with LOINC (Logical Observation Identifier Names and Codes) [5]. This is also the choice made by other hospitals showing a more or less complete interfacing with LOINC [4]. With this regard, the biomedical analysis dictionary is a perfect example of local terminology interfaced with a reference terminology [3]. This dictionary is also linked to adjunct data such as the list of hospital user facilities as well as their contacts.

In this project, we aim to implement an editing and maintenance platform of the AP-HP biomedical analysis dictionary that may please the requirements related to both the repository and maintenance and diffusion processes. The current repository management software (spreadsheet) shows some adaptation limitations to the dictionary requirements and perspectives. To achieve this goal, our efforts focus on an ontology model definition which supports the representation of the terminologies used and also of adjunct knowledge. This modeling effort must permit the multi-terminological representation, terminology update (e.g. half-yearly LOINC

[1] Assistance Publique – Hôpitaux de Paris. Public hospital of Paris and its suburbs. AP-HP is the largest hospital system in Europe and one of the largest in the world.

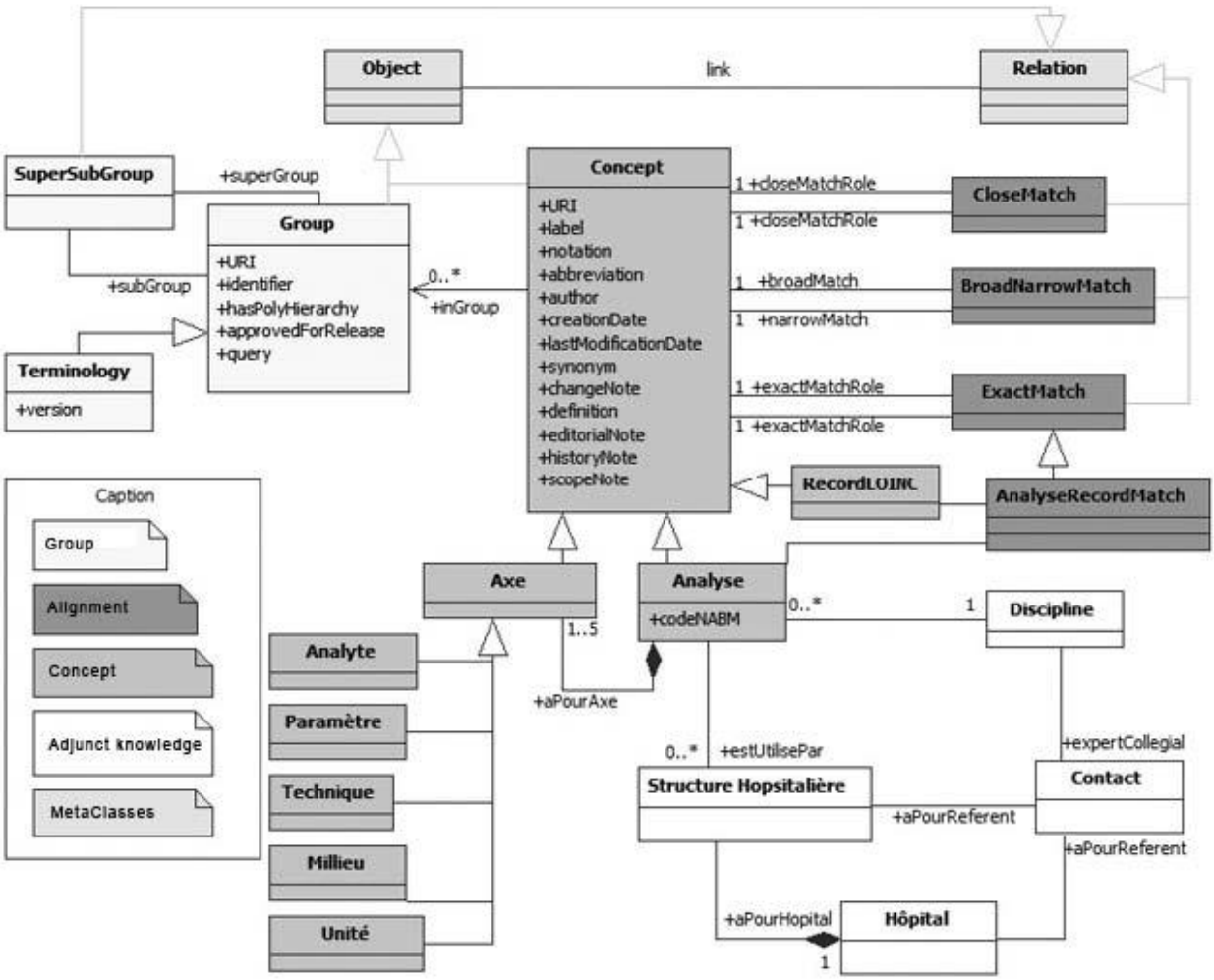

Figure 1. Simplified model of UML class of the AnaBio Ontology model.

update), and possibility to store the translation of LOINC. The data formalization must also allow improving data quality.

2 Method and Results

Our method apprehends the diversity of the terminologies pattern expressivity by defining a unique ontology representation model. This unique model presents the advantage to integrate the different terminologies within a single server and thus, to allow the editing of these repositories. The new platform implementation requires the transmission of semi-structured data to a structured model representation which relies on the knowledge engineering techniques [1]. This change implies modeling, data migration, and integration/validation steps.

The modeling task is conducted in close collaboration with the Terminology Maintenance Unit. This collaboration aims understanding the usefulness of each component and apprehending the new needs which impact the model to design. For example, the addition of status properties allows a better traceability of the components over the time. The structuring of the spreadsheet existing in tabs and columns constitutes a first organization step, discipline understanding, and data exchange need integration. The model is enough generic to represent any type of terminology including LOINC, AnaBio and also future resources that will be useful to improve the interoperability, such as SNOMED-CT. However, this model remains extensible to consider the particularities of each terminology (for example the NABM codes for biomedical

analysis results) but also to link the AnaBio dictionary to the adjunct knowledge such as hospital facilities, contacts, etc. Our approach does not pretend to define an ex nihilo model but wants to be a good practice paradigm for the controlled vocabulary representation. Our method uses and extends parts of modeling in existing norms and standards such as SKOS (Simple Knowledge Organization System) [6] and BS 8723 (British Standard 8723) [2]. The OWL language and its expressivity in description logic are used to describe our model [7] which is presented in Figure 1. Parallel to the model design, the task of **data migration** begins. This requires transforming the entire spreadsheet data to allow their integration and the conformity alignment with the new formal model. Modeling and data migration stages allow an iterative refinement work. To be validated, modeling suggestions are imported into the platform with the migrated data. During the **validation** task, the terminology maintenance team validate, correct and improve the ontology. After 6 validation cycles, the platform deployment in the production environment intervenes.

The improvement in data quality included in the AnaBio dictionary is a major point of the results obtained with this project. The transition from semi-structured data (spreadsheet) to structured data (by the formal model) has forced the correction of data considered as incoherent. Most of these inconsistencies were differences of breakage, spelling or absence of normalization of a value which should be identical. For example, “cysterceques”, “Cysterceque” and “Cysticerques anticorps” will be change to “Cysticerques anticorps”. These corrections aim to improve data quality.

3 Conclusion

This ontology model is a particularly suitable and scalable solution for the needs of a terminology used daily. Contrarily to a XLS file which is constraint by its structure, this model can be extended without impacting the controls, exports and statistics already implemented. Its generic nature allows the future integration of other terminologies. It also allows the definition of restriction, inference and control rules through its formal definition. The platform implementation to manage biomedical analyses and their associated data highlights some issues which were hidden until then. More than 10% of original data have been corrected during this project. The implemented solution automates and integrates a large number of tasks (automatic index creation, input constraint control defined in the model, etc.), releasing the team in charge of the AnaBio dictionary from proceedings outside of their expertise.

References

1. AussenacGilles, N.: Méthodes ascendantes pour l’ingénierie des connaissances. Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse, France (décembre 2005), ftp://ftp.irit.fr/IRIT/CSC/HDR-Aussenac13fev-06.pdf
2. BS8723: Structured vocabularies for information retrieval, part 4: Interoperability between vocabularies, (2008)
3. Daniel, C., Buemi, A., Mazuel, L., Ouagne, D., Charlet, J.: Functional requirements of terminology services for coupling interface terminologies to reference terminologies. In: Studies in health technology and informatics. vol. 150, p. 205 (2009)
4. Lin, M., Vreeman, D., McDonald, C., Huf, S.: A characterization of local loinc mapping for laboratory tests in three large institutions. Methods Inf Med 2010, 49 (2009)
5. McDonald, C., Huf, S., Suico, J., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook, J., et al.: Loinc, a universal standard for identifying laboratory observations: a 5-year update. Clinical chemistry 49(4), 624 (2003)
6. Miles, A.: Skos: requirements for standardization. In: DC-2006: Proceedings of the International Conference on Dublin Core and Metadata Aplications. pp. 55 64 (2006)
7. Vandenbussche, P.Y., Charlet, J.: Méta-modèle général de description de ressources terminologiques et ontologiques. In: Ingénierie de la Connaissance (IC) (2009)