International Journal of Medical Informatics (2003) **70**, 317–328

ELSEVIER

International Journal of Medical Informatics

www.elsevier.com/locate/ijmedinf

Developing the ontological foundations of a terminological system for end-stage diseases, organ failure, dialysis and transplantation

Christian Jacquelinet[a,*], Anita Burgun[b], Denis Delamarre[b], Nigel Strang[a], Sami Djabbour[a], Bernard Boutin[a], Pierre Le Beux[b]

[a] *Département Medical et Scientifique, Etablissement français des Greffes, 5, rue Lacuée, Paris 75012, France*
[b] *Laboratoire d'informatique Médicale, Faculté de Médecine, Rennes, France*

Received 16 December 2002; accepted 28 February 2003

KEYWORDS
Organ failure;
Dialysis;
Transplantation;
Thesaurus;
Terminological system;
Semantic interoperability;
EDI;
Ontology;
Natural language processing;
Conceptual graphs;
Semiotics

Summary The Etablissement français des Greffes (EfG) is a state agency dealing with Public Health issues related to organ, tissue and cell transplantation in France. The evaluation of organ retrieval and transplantation activities, one of its missions, is supported by a national information system (EfG-IS). The EfG-IS is moving towards a new n-tier architecture comprising a terminology server for end-stage diseases, organ failure, dialysis and transplantation (EfG-TS). Following a preliminary audit of the existing coding system and in order to facilitate data recording, to improve the quality of information, to assume compatibility with terminological existing standards and to allow semantic interoperability with other local, national or international registries, a specific work has been conducted on the thesauri to integrate within the EfG-TS. In this paper focusing on the server's content rather than the container, we report first the functional and cognitive requirements that resulted from the preliminary audit. We then describe the methodological approach used to build the terminological server on ''sound ontological foundations''. We performed the semantic analysis of existing medical terms to set up disease description frame-like structures. These diseases description frames consist of a limited set of nosological discriminating slots such as etiology, semiology, pathology, evolution and associated diseases. Each relevant medical term is thus associated to a concept defined and inserted within a hierarchy according to disease description frame resulting from the semantic analysis. Last, because this terminological server is shared by various transplant and dialysis centers to record patient data at different time point, contextualization of terms appeared as one of the functional requirements. We will also point out various contexts for medical terms and how they have been taken into account.

1. Introduction

The Etablissement français des Greffes (EfG) is a state agency in charge of public health issues

*Corresponding author.
E-mail address: christian.jacquelinet@univ-rennes1.fr (C. Jacquelinet).

1386-5056/03/$ - see front matter
doi:10.1016/S1386-5056(03)00046-7

related to organ, tissue and cell transplantation in France. According to the law 93-43 of January 1994 related to Public Health and Social Care, EfG is responsible for ''the registration of patients on the national waiting list, the management of this list and the allocation of all organs, retrieved in France or outdoors''. Ministerial order 94-870 of October 1994 gave to EfG the charge of ''the good application of the rules related to the management of the national waiting list, to the distribution of the cadaveric organs and to the evaluation of retrieval and transplantation activities by organ and by transplantation center'' [1]. To fulfill these missions, EfG maintains a national information system (EfG-IS) in which the transplantation teams record patients data at the registration on the national waiting list, at the time of transplantation if any and during the follow-up before and after transplantation. Furthermore, EfG is now in charge of the Renal Epidemiology and Information Network (REIN) devoted to the follow-up of all end-stage renal disease patients treated by dialysis or transplantation [2], reconnecting the evaluation of kidney transplantation to the evaluation of dialysis.

In order to facilitate data recording, to improve quality of data, to allow collaboration between local, national and international registries, the question of the global architecture of the EfG-IS and the question of its coding system have been raised. Such issues led us to consider how to organize the cooperation between a set of production databases and an exploitation-oriented data-warehouse within a multi-source information system [3,4], raising again the central question of the EfG-IS coding system in terms of Electronic Data Interchange (EDI), standardization and semantic interoperability. Semantic interoperability is defined as the capability of information systems to use the information that they exchange, to deal with the various ways to represent the real world entities, with the differences in the meaning or in the uses of data, and in the relations between objects. Semantic heterogeneity concerns variations in data structure (field names, attributes, and relations), in covered domains, in underlying terminology and in exploitation goals [5,6]. The standardization of the terms and thesauri has longstanding been recognized as a first step toward semantic interoperability [7,8]. AFNOR in France and CEN Technical Committee 251 (TC251) in Europe have working groups preparing standards in the scope of healthcare or Health Level Seven project (HL7) in the US [6,9]. But all these works are mostly dealing with format or more general medical information [10].

The existing EfG-IS—a RTC based client-server classical architecture, is actually moving toward a secured web-based n-tier modular architecture comprising a terminological server. This terminological server is the container. Our aim here is to deal first with the content and to develop a terminological server for end-stage diseases, organ failure, dialysis and transplantation on sound ontological foundations. In this paper, we report first the functional and cognitive requirements that resulted from the preliminary audit. We then describe the methodological approach used to build the ontological foundations of the server with the semantic analysis of existing medical terms. We report the resulting representational models used to describe diseases and how we use them to insert within a hierarchy of concepts relevant medical terms.

2. Materials and methods

A preliminary audit was performed including a quantitative analysis of the existing EfG coding system and a qualitative analysis checking completeness, consistency, ambiguity and implicitness of terms. This audit showed that the thesaurus was physically split into a set of unconnected tables of values related to: (i) the registration of the patient on the waiting list for the initial disease, (ii) the follow-up of the patient before and after the transplantation for the complications, (iii) the follow-up of the patient before and after the transplantation for the causes of death. The thesaurus consisted of lists of medical terms, realizing a catalog without hierarchy, explicit differentiation principles nor compositional principles. It was thus poorly structured, except the separation of terms according to the roles of diseases (indication, complication, cause of death) across the healthcare history of the patient. The terms themselves were not referring to medical standards, except terms used for kidney diseases inspired from those of the ERA-EDTA register, although not a standard [11]. The lists of available terms were lacking of completeness. Some terms were poorly appropriate and were not used as demonstrated by the quantitative study. The qualitative analysis showed that 25% of the terms proposed for initial diseases had at least one defect among the following: ambiguity, incompleteness, implicit or inconsistency. Duplicates and overlaps were noticed such as: *Acute Hepatic Insufficiency* and *Hyper-acute or Fulminant Hepatitis*. For the complications, the lack of pertinence of the lists of terms was worsened by the absence of contextua-

lisation with the type of organ: for example, a *stenosis of the artery of the renal graft* is evidently not pertinent in the context of lung transplantation.

2.1. Defining functional and cognitive requirements

Results from the preliminary study attested of potential difficulties for the epidemiological exploitation of the data due to structural anomalies of the existing thesaurus. Rules of literal formulation of terms and a knowledge representational model appeared necessary to define according to the following cognitive and ontological requirements: (1) double disambiguation of terms provided by their insertion in a hierarchy and by the possibility to access to inclusion criteria making possible to decide if an individual belongs to the corresponding class; (2) no a priori intent of parsimony in the number of terms, in order to avoid an incomplete coverage of the domain; (3) coherence with the existing terminological sources and standards; (4) use of a knowledge representational model having the capability to support both the interoperability requirements and epidemiological exploitation goals; (5) description of terms according to schema able to represent the temporal dimension of the events before and after transplantation and able to support the chaining of disease's causality such as: hepatitis/cirrhosis/liver cancer. The granularity of information in the existing thesaurus was heterogeneous, ranging from a dozen terms for heart diseases to more than 60 detailed terms for renal diseases. The reunification of terms within the terminological system appeared as a need. To facilitate the selection of terms, their insertion within a hierarchy emerged as a requirement. Nevertheless, a strong evidence was that the granularity of terms should not be identical all along the patient's history nor it should be the same for all users. Very precise terms were required to code the initial disease at the registration on the waiting list, which is performed by organ specialists (nephrologists, hepatologists, cardiologists, pneumologists) whereas more general terms were desired to code co-morbidity or complications. The EfG's terminological server must be able to be used within the framework of the current EfG-IS and the REIN-IS. It will be shared by various kind of transplant teams and dialysis centers physicians to record patients data at different time points. Thus, contextualisation of terms appeared as another functional requirement.

2.2. Terminological resources and test bed

Organ failure and transplantation is a specific medical knowledge domain. This domain may be divided according to the organ into sub-domains that overlap with more classical medical domains related to the same field (heart transplantation and cardiology, for example) or with infectious diseases or malignancies for complications. A specific terminology is thus required only for a restricted subset of terms used to describe some graft or immuno-suppressive therapies related events. It is thus possible to check existing medical terminologies such as ICD-10 families [12–14] and medical terminological systems such as SNOMED-RT [15,16], as well as thesauri related to international registries for organ failure or transplantation because of the specificity of the domain to cover and because of our intention to deal with EDI. Terms from the EfG ad-hoc thesaurus are also to study because of the necessity to recover old records. Our problematic is then to integrate subsets of terms coming from various terminological systems within the same support. This problem is very close to the one of the UMLS [17,18], transposed in a smaller scope but thinner grain sub-domain. Terms coming from the following resources have been taken into account: (i) international registries for transplantation, such as the International Society for Heart and Lung Transplantation registry, the European Liver Transplant Registry, the European Renal Association and European Dialysis and Transplantation registry; (ii) ICD-10, UMLS concepts and relations linked to transplantation; (iii) French accredited specialized thesauri such as the thesaurus of the French Society for Renal Diseases, the thesaurus of the French Society for Hepato-Gastroenterology. A first test bed of medical terms used to describe diseases related to initial disease leading to the registration on the waiting list (10), complications (5) and causes of death (10) was selected to build underlying diseases knowledge schemes, if possible the most general and sharable ones. A second test bed of 100 medical terms randomly selected was used to check the robustness of these models and served within the early evaluation of the terminological server.

2.3. Knowledge representational model and disease description frames

A major issue in this study was to work out a knowledge representational model able to support the ontology of the domain independently of the final implementation. According to [26], the ontol-

[**Disease**: {*cardiopathy, disease, glomerulonephritis, myocarditis, nephropathy, nephritis*}: #234]
[(**has_a**): {*$*}: #5631], **signature**(#5631)={(#234, #1359),…}
[**Lesion**: {*cardiopathy, disease, glomerulonephritis, myocarditis, nephropathy, nephritis*}: #1359]
[(**which_is**): {*$*}: #764], **signature**(#764)={(#1359, #75563),…}
[**Inflammation**: {*glomerulonephritis, myocarditis, nephritis*}: #75563]
[(**has_a**): {*$*}: #884], **signature**(#884)={(#1359, #2318),…}
[**Location**: {*cardiopathy, disease, glomerulonephritis, nephropathy, nephritis*}: #2318]
[(**which_is**): {*$*}: #972], **signature**(#972)={(#2318, #769), (#2318, #621), …}
[(**which_is**): {_}: #7641], **signature**(#7641)={(#2318, #7095), (#2318, #99674), …}
[**Heart**: {*heart*}: #7095]
[**Heart**: {*cardiopathy, myocarditis*}: #769]
[**Kidney**: {*renal*}: #99674]
[**Kidney**: {*nephropathy, nephritis, glomerulonephritis*}: #621]
[(**has_a**): {*$*}: #2648], **signature**(#2648)={(#621, #8563), (#769, #8563) …}
[Part: {*glomerulonephritis, myocarditis*}: #8563]
[(**which_is**): {*$*}: #2284], **signature**(#2284)={(#8563, #55327),…}
[**Glomerulus**: {*glomerulonephritis*}: #55327]
[**Myocardium**: {*myocarditis*}: #55327]
[(**has_a**): {*$*}: #5631], **signature**(#5631)={(#234, #22144),…}
[**Etiological Process**: {*cardiopathy, disease, glomerulonephritis, myocarditis, nephropathy, nephritis*}: #22144]
[(**which_is**): {_}: #53398], **signature**(#53398)={(#22144, #11209),…}
[**Ischemia**: {*ischemic*}: #11209]

Fig. 1 A sample of the KB defining a semiotic network by linking lexicalized concepts to others by the mean of lexicalized conceptual relations and their signatures.

ogy involves the concepts pertaining to the domain of transplantation, dialysis and end-stage diseases and the relations between these concepts. The Conceptual Graphs (CGs) formalism was selected as underlying knowledge representational model on the following grounds. CGs are based on Pierce existential graphs and on semantic networks [19,20]. If one consider their readability and their ability to support Natural Language, CGs provide an interesting knowledge representational model. In the medical domain, CGs have been shown relevant for the representational and ontological foundation of terminological systems for anatomy, disease process, treatments and healthcare procedure [21–25]. Concept type lattice and Conceptual Relation hierarchy can support the ontology of the domain. Type definition, canonical graphs, schemata and prototypes provide supports for a knowledge base. Furthermore, the CG model provides operations to combine CGs and to derive CGs from other CGs. Specialization, generalization operations, graph projection, join and inference rules constitute a rich knowledge processing toolbox that might be useful for Knowledge Processing. In his recent work [27], J. Sowa embodies the CG model in the field of Knowledge Representation and Ontology building. To represent objects in a simple and stereotyped manner, it was decided to use frame-like structures. One of the interests of frames is to formulate what is awaited, so that implicit information masked in some terms can be more easily detected. Description frame is also a surface representation easily readable by an expert clinician by guiding it in his description. At least partial, an automated completion of the frames was required.

2.4. Principles of semantic analysis of medical terms with RIBOSOME

A knowledge extraction tool called RIBOSOME was used to perform semantic analysis of the terms selected in test beds [28]. RIBOSOME comprises a knowledge base (KB) and a parser that turns short texts and medical terms into CGs. The KB integrates within the same support the concept type lattice, the conceptual relations hierarchy and the conceptual relations signatures (i.e. valid head and tail concepts for a given conceptual relation). Furthermore, the KB supports an extension to CG model: concepts and conceptual relations are associated to lexia that can be words, part of words or locutions. Thus, the KB also integrates a semantic lexicon relating lexical entries to their relevant conceptual structures. Even the blank space between words (_) is a valid lexical entry. The basic structure of the KB is referred to as a *cognitive sign*. It associates a type label, a set of lexia and a contextual marker #i:

[**concept_type_label**$_a$: {*lexia*}$_a$: #i].

In the CG model, CR signatures are defined at the semantic level, with no formal link to the semantic lexicon. A key feature with RIBOSOME is the specification of conceptual relations signatures at the semiotic level: lexicalized relational cognitive signs are associated to head and tail lexicalized conceptual cognitive signs:

[**concept_type_label**$_a$: {*lexia*}$_a$: #i] → [(**CR_type_label**): {*lexia*}$_b$: #j] → [**concept_type_label**$_c$: {*lexia*}$_c$: #k]

where, the couple (#i, #k) is one of the **signatures** of #j.

Cognitive signs are thus inserted into a semantic network alternating lexicalized conceptual and relational cognitive signs. As a consequence: (i) the resulting semantic network is referred as to a *semiotic network*; (ii) meta-knowledge associated with conceptual relations signatures includes the lexical context related to the lexical entries of refining cognitive signs; (iii) semantic constraints defined for conceptual relations are associated to more precise lexical constraints that rules the composition of cognitive signs during parsing of texts. A given lexical referent may be shared by many cognitive signs, allowing the treatment of generalized polysemy. A special *lexical entry*, noted $, is used for conceptual relations to restrict during parsing the composition of head and tail cognitive signs only to those sharing the same lexical referent. For example, let's consider the KB in Fig. 1 linking Conceptual and Relational cognitive signs with relation's signature.

With such a KB, we can illustrate the resulting semantic analysis (SemAnalysis) for some terms:

(1) **SemAnalysis**("*disease*") =

[**Disease**: *disease*: #234]-
[(**has_a**): $: #5631] → [**Lesion**: *disease*: #1359]-
[(**has_a**): $: #884] → [**Location**: *disease*: #2318]
[(**has_a**): $: #5631] → [**Etiological Process**: *disease*: #22144]

Such a semiotic scheme considers that the word *disease* denotes intrinsequely a **Disease** with an unspecified **Etiological Process** and an unspecified **Lesion**, which has an unspecified **Location**: the **Lesion** exists, its type is not specified, it has a **Location** whose type is unspecified. With such an approach, the denotation of *disease* is not limited to the atomic concept **Disease** but related to a molecule of meanings.

(2) **SemAnalysis**("*heart disease*") =

[**Disease**: *disease*: #234]-
[(**has_a**): $: #5631] → [**Lesion**: *disease*: #1359]-
[(**has_a**): $: #884] → [**Location**: *disease*: #2318]-
[(**which_is**): _: #7641] → [**Heart**: *heart*: #7095]
[(**has_a**): $: #5631] → [**Etiological Process**: *disease*: #22144]

By composition, *heart* specifies the **Location** of the **Lesion** given by the semiotic scheme of *disease*. The white space "_" is the lexical referent in [(**which_is**): _: #7641]. The lesion exists, its type is unspecified, its location is specified by composition.

(3) **SemAnalysis**("*cardiopathy*") =

[**Disease**: *cardiopathy*: #234]-
[(**has_a**): $: #5631] → [**Lesion**: *cardiopathy*: #1359]-
[(**has_a**): $: #884] → [**Location**: *cardiopathy*: #2318]-
[(**which_is**): $: #972] → [**Heart**: *cardiopathy*: #769]
[(**has_a**): $: #5631] → [**Etiological Process**: *cardiopathy*: #22144]

The location of the lesion is intrinsequely given by *cardiopathy*. The auto-combining symbol "$" is the lexical entry in [(**which_is**): $: #972]. The word *cardiopathy* simultaneously triggers five concepts that auto-combine to each other during the parsing.

(4) **SemAnalysis**("*myocarditis*") =

[**Disease**: *myocarditis*: #234]-
[(**has_a**): $: #5631] → [**Lesion**: *myocarditis*: #1359]-
[(**which_is**): $: #764] → [**Inflammation**: *myocarditis*: #75563]
[(**has_a**): $: #884] → [**Location**: *myocarditis*: #2318]-
[(**which_is**): $: #972] → [**Heart**: *myocarditis*: #769]-
[(**has_a**): $: #2648] → [**Part**: *myocarditis*: #8563]-
[(**which_is**): $: #2284] → [**Myocardium**: *myocarditis*: #55327]
[(**has_a**): $: #5631] → [**Etiological Process**: *myocarditis*: #22144]

The location and the type of the lesion (some inflammation) is intrinsequely given by *myocarditis*. Furthermore, the word *myocarditis* denotes **Heart**, which has a **Part**, which is the **Myocardium**.

(5) **SemAnalysis**("*ischemic disease*") =

[**Disease**: *disease*: #234]-
[(**has_a**): $: #5631] → [**Lesion**: *disease*: #1359]-
[(**has_a**): $: #884] → [**Location**: *disease*: #2318]
[(**has_a**): $: #5631] → [**Etiological Process**: *disease*: #22144]-
[(**which_is**): _: #53398] → [**Ischemia**: *ischemic*: #11209]

These examples illustrate the principles of lexical semantics existing in RIBOSOME. Larger amount of words can share the same conceptual structures, so that lexical and semantic knowledge acquired to perform the semantic parsing of some terms improve the compositional possibilities with other previously analyzed terms. One also point out that homogeneity of rules of lexical semantics is an important help for the standardization of

knowledge representation and offers huge possibilities for the reengineering of existing terminologies.

2.5. Building RIBOSOME's semantic lexicon

The improvement of the semantic lexicon of RIBOSOME relies on a cyclic process comprising.
(1) The initiation of a targeted knowledge representational scheme using a CG candidate to represent the semantic of the term; for *acute inflammatory myocardial heart disease* we could have the following targeted CG:

[Disease]-
(has_a) → [Lesion]-
(which_is) → [Inflammation]
(has_a) → [Location]-
(which_is) → [Heart]-
(has_a) → [Part]-
(which_is) → [Myocardium]
(has_a) → [Evolutivity]-
(which_is) → [Acuteness]

(2) each component of the targeted CG is then lexicalized, leading to a semiotic scheme; for example:

[**Disease**: *disease*]-
[(**has_a**): $] → [**Lesion**: *disease*]-
[(**which_is**): _] → [**Inflammation**: *inflammatory*]
[(**has_a**): $] → [**Location**: *disease*]-
[(**which_is**): _] → [**Heart**: *heart*]-
[(**has_a**): _] → [**Part:** *myocardial*]-
[(**which_is**): $] → [**Myocardium**: *myocardial*]
[(**has_a**): $] → [**Evolutivity**: *disease*]-
(**which_is**) → [**Acuteness**: *acute*]

(3) If not previously existing, then relevant concepts and conceptual relations are created as subtypes of existing concepts or CRs within the KB: for example, [**Heart**] is created as a subtype of a pre-existing [**Organ**], its lexical referent are recorded by the user; the system returns a contextual marker.
(4) Robustness and coherence of semiotic scheme is then assessed by reanalyzing previously analyzed terms, checking for noise due to inappropriate compositions and for silence due to inexistence or disappearance of appropriate compositions.

The semantic analysis of a medical term with RIBOSOME is illustrated in Fig. 2. The input is firstly segmented into valid lexical entries: word, locution or entire label of a disease. Words composing locutions and disease labels are analyzed at the same time. The next step is the contextual selection of conceptual structures, triggered by the lexical input and their combination in more complex primary conceptual structures. A set of post-processing functions is available: among them, one joins sub-graphs of the primary output into a frame like format close to the standard conceptual graph linear format (CGLF); another turns the description frames into more readable format where conceptual relation labels are explicit, so that medical experts working with our group easily understand the graphs. CG's are generated for medical terms and support the semantic structure of the terms. Conceptual structures are inserted into a semantic network that is the support of the domain ontology and that permits the re-use of the acquired semantic knowledge in next analysis.

3. Results

3.1. Terms and diseases description frames

The comparison of the semantic structures generated for the test bed allowed us to propose a general scheme for the description of a disease. One prominent result is that the formulation of terms can be specified according to a limited set of nosological discriminating slots such as etiology, semiology, pathology, evolution and associated diseases. Generic terms, as well as ambiguous or implicit terms, appeared as unspecified terms according to some slots. Such a description frame supports the type definition of a term: the head concept is the genus whereas discriminating nosological slots act as differentiae. Not all clinical signs that can be encountered in a disease need to be described, nor all lesions, associated diseases or etiological processes, only those that constitute a sufficient condition. The description frame [Disease] is used to define a disease responding to a given term. It embeds others description frames such as [Etiological Process], [Finding], [Lesion] or even [Disease] again for secondary or ternary diseases. Ambiguous and implicit terms were under-specified at the lexical level (in the formulation of the term) but could be improved by the expert with the completion of unspecified slots.

[Disease]
(label) → [Term]
(code) → [Code]
(discriminating etiological process) → [Etiological Process]
(discriminating clinical finding) → [Finding]
(discriminating pathological lesion) → [Lesion]
(discriminating associated disease) → [Disease]

[Etiological Process]-
(label) → [Term]
(whose type is) → [Process_type]
(discriminating agent) → [Agent]
(discriminating patient) → [Patient]
[Lesion]
(label) → [Term]
(whose type is) → [Leison_type]
(discriminating location) → [Anatomical Component]
(discriminating evolution characteristic) → [Evolution]

A description frame can also be specialized if its specialization is often reused: [Liver Disease], [Tumoral Disease] for example.

[Liver Disease]-
(label) → [Term]
(code) → [Code]
(discriminating etiological process) → [Etiological Process]
(discriminating clinical finding) → [Finding]
(discriminating pathological lesion) → [Lesion]-
(discriminating location) → [Liver]
(discriminating associated disease) → [Disease]

[Tumoral Disease]-
(label) → [Term]
(code) → [Code]
(discriminating etiological process) → [Etiological Process]
(discriminating clinical finding) → [Finding]
(discriminating pathological lesion) → [Tumor]
(discriminating associated disease) → [Disease]

3.2. Semantic integration, hierarchy and granularity

One interesting rule related to conceptual refining is that a partially specified graph G1 subsumes a more specified graph G2:

let G: [C1] → (RC1) → [C2], G′: [C1] → (RC1) → [C3], if C2 > C3 then G > G′.

As a corollary, because [C] > [C] → (RC) → [C′], a "more detailed" CG G″: [C1] → (RC1) → [C3] → (RC2) → [C4] is such that G″ < G′ < G.

These rules provide an efficient tool for the organization of the hierarchy on formal ontological foundations. As an example, let us consider the following description frames for Wilson's disease related Hepatopathies in two clinical presentations:

[Liver Disease]-
(label) → [**Wilson's disease Fulminant Hepatitis**]
(code) → [EfG#x]
(discriminating etiological process) → [Unspecified]
(discriminating clinical finding) → [Unspecified]
(discriminating pathological lesion) → [**Necrosis**]-
(discriminating evolution) → [**Hyper-Acuteness**]
(discriminating location) → [Liver]
(discriminating associated disease) → [**Wilson's Disease**]

[Liver Disease]-
(label) → [**Wilson's disease Chronic Hepatitis**]
(code) → [EfG#x]
(discriminating etiological process) → [Unspecified]
(discriminating clinical finding) → [Unspecified]
(discriminating pathological lesion) → [**Hepatitis**]-
(discriminating evolution) → [**Chronicity**]
(discriminating location) → [Liver]
(discriminating associated disease) → [**Wilson's Disease**]

Both graphs have a common supertype that denotes Wilson's disease Hepatopathy with unspecified lesions. With the same approach, the partial specification of the discriminating associated disease leads to the description frame of all metabolic liver diseases.

[Liver Disease]
(label) → [**Wilson's disease Hepatopathy**]
(code) → [EfG#z]
(discriminating etiological process) → [Unspecified]
(discriminating clinical finding) → [Unspecified]
(discriminating pathological lesion) → [**Lesion**]-
(discriminating evolution) → [**Unspecified**]
(discriminating location) → [Liver]
(discriminating associated disease) → [**Wilson's Disease**]

[Liver Disease]-
(label) → [**Metabolic Liver Disease**]
(code) → [EfG#w]
(discriminating etiological process) → [Unspecified]
(discriminating clinical finding) → [Unspecified]
(discriminating pathological lesion) → [Lesion]-
(discriminating evolution) → [Unspecified]
(discriminating location) → [Liver]
(discriminating associated disease) → [**Metabolic Disease**]

With the same approach, the non-specification of both copper concept and discriminating associated disease is leading to the description frame of all metabolic diseases. It defines Wilson's disease as a subtype of metabolic disease. One can see here also that Wilson's disease fulminant hepatitis and Wilson's disease Hepatopathy are not hyponyms of Metabolic Disease but an hyponym of Metabolic Liver Disease.

[Metabolic Disease]-
(label) → [**Wilson's disease**]
(code) → [EfG#k]
(discriminating etiological process) → [Metabolic Disorder]-
(discriminating patient) → [**Copper**]
(discriminating clinical finding) → [Unspecified]
(discriminating pathological lesion) → [Unspecified]
(discriminating associated disease) → [Unspecified]

[Disease]-
(label) → [**Metabolic disease**]
(code) → [EfG#j]
(discriminating etiological process) → [**Metabolic Disorder**]-
(discriminating patient) → [**Unspecified**]
(discriminating clinical finding) → [Unspecified]
(discriminating pathological lesion) → [Unspecified]
(discriminating associated disease) → [Unspecified]

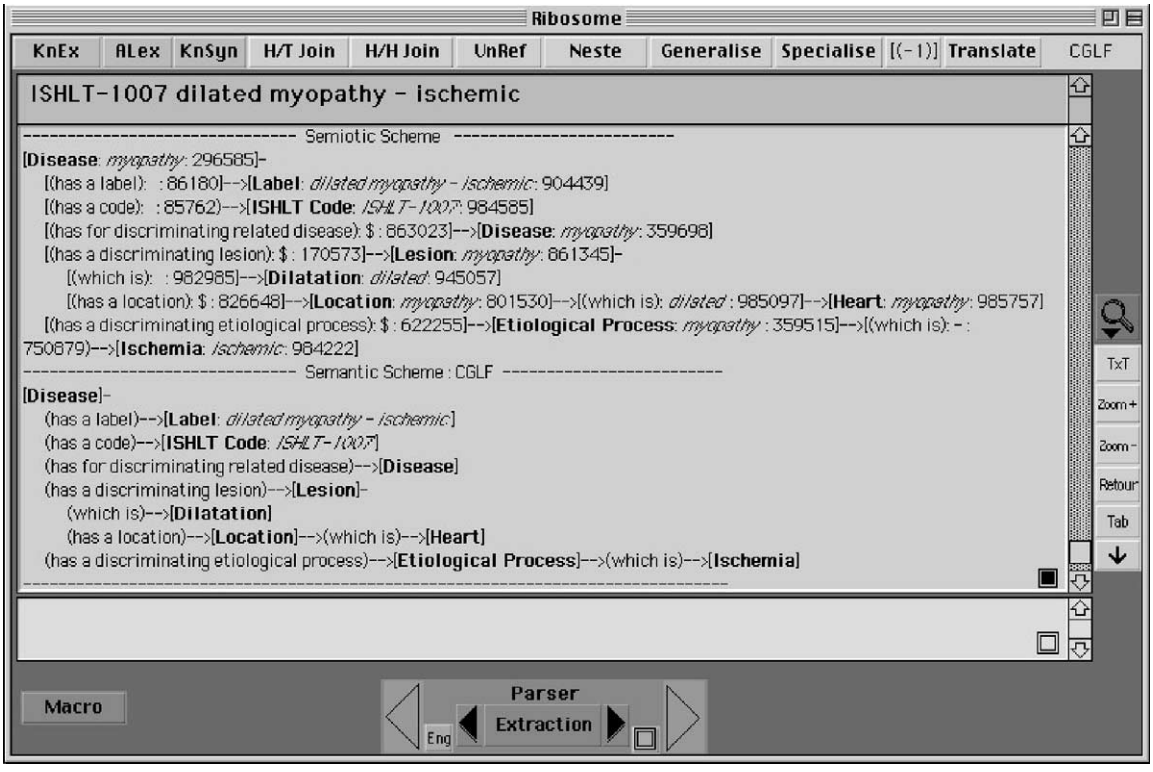

Fig. 2 Semantic Analysis of an ambiguous term existing in the ISHLT thesaurus with RIBOSOME. The disambiguation of *myopathy* (the complete term should be cardiomyopathy) is given, in this semiotic scheme, by the adjective *dilated* featuring in the co-text and used as a lexia for the conceptual relation (which_is) that links a special local context #985757 of [**Heart**] with the widely shared context #801530 of [**Location**].

[**Disease**]-
(label)→[**Term**]
(code)→[**Code**]
(hyperhonyms)→{[**Disease$_1$**], [**Disease$_2$**], …}
(hyponyms)→{[**Disease$_3$**], [**Disease$_4$**], [**Disease$_5$**], [**Disease$_6$**],…}
(discriminating pathological lesion)→[**Lesion**]
(discriminating etiological process)→[**Etiological Process**]
(discriminating clinical sign)→[**Sign**]
(discriminating associated disease)→[**Disease$_7$**]
(information production context)→{[**Temporal Role**], [**Treatment Procedure**], [**Communication Situation**]}
(information exploitation context)→{[**Temporal Role**], [**Treatment Procedure**], [**Exploitation Situation**]}

Fig. 3 Local contextual slots are added to disease description frames.

3.3. Taking into account information production contexts

One interesting result of our study is the definition of a set of information production contexts and a set of information exploitation contexts. An information production context is defined by the combination of (i) the temporal roles of the disease (indication/initial disease, co-morbidity, complication, cause of death) (ii) the healthcare and treatment procedure: in our case dialysis and kidney, heart, liver, lung, pancreas or intestinal transplantation and (iii) the communication situation (patient data record, EDI, browsing through the hierarchy, user's interface grouping item). This approach enables for example a given medical term to be used by heart transplant teams as complication to record patient data, whereas kidney transplant teams use it as an indication or a co-morbidity item only to browse within the hierarchy of terms. A grouping item is devoted to dynamically drive the construction of user's interface screens with check boxes. The users will systematically check the presence or the absence of a grouping item. The absence of a grouping item invalidates the presence of its hyponyms according to the hierarchy.

An information exploitation context is defined by the combination of (i) the temporal roles of the disease (indication/initial disease, co-morbidity, complication, cause of death) (ii) the healthcare and treatment procedure (dialysis, kidney, heart, liver, lung, pancreas or intestinal transplantation) and (iii) the statistical exploitation situation (grouping item for counting, risk factor grouping item, grouping item for event free survival). The contexts of production and exploitation of the information are summarized in Fig. 3 by adding contextual slots to disease description frames. A given disease may have many information production and exploitation contexts. Contexts are local slots whereas discriminating nosological slots are inheritable.

The treatment procedure is related to the kind of users. It constitutes the user's part of the context (who is using the term?). Interestingly, a given disease as a complication or as an indication is not exactly the same object. It constitutes the temporal part of the context (what kind of temporal role for the term?). Last, the communication situation denotes the use of the medical term.

In Fig. 4, we can read that a *Glomerular nephropathy* will be usable as an indication by Kidney transplant teams to browse within the hierarchy of diseases (a more precise hyponym is required to record patients data). It will be valid as co-morbidity for other transplant teams that will also have to check systematically its presence or its absence as a potential complication because this disease was labeled as a user's interface grouping item. This disease will also be a grouping item to count *Glomerular nephropathy* and all its hyponyms as an indication of transplantation.

4. Discussion–conclusion

In respect to the cognitive requirements that resulted from the audit, we used terms coming from existing terminologies to set up our domain ontology. The use of a knowledge extraction tool appeared as an important help to compare the semantic structures generated for medical terms and allowed us to propose a general schema for the description of a disease. This work showed us that the ontology to build for end-stage disease, organ failure, dialysis and transplantation was nothing else than a nosology where the pathological domain is to split into distinct medical entities and that it was possible to model diseases using description frames with a limited and stereotyped set of discriminating nosologic slots. Linking med-

[**Nephropathy**]-
(label)→[**Glomerular Nephropathy**]
(code)→[**EfG#659**]
(hyperhonyms)→[**Nephropathy**]
(hyponyms)→{[**Primary Glomerular Nephropathy**], [**Secondary Glomerular Nephropathy**], …}
(discriminating pathological lesion)→[**Lesion**]-
(which is)→[**Unspecified**]
(has a location)→[**Location**]→(which is)→[**Glomerulus**]
(discriminating etiological process)→[**Etiological Process**]→(which is)→[**Unspecified**]
(discriminating clinical sign)→[**Finding**]→(which is)→[**Unspecified**]
(discriminating associated disease)→[**Disease**]→(which is)→[**Unspecified**]
(information production context)→{[**Indication**], [**Kidney Transplantation**], [**Browsing**]}
(information production context)→{[**Co-morbidity**], [**Heart, Liver Transplantation**], [**Data Recording Item**]}
(information production context)→{[**Complication**], [**Heart, Liver Transplantation**], [**User's Interface Grouping Item**]}
(information exploitation context)→{[**Indication**], [**All Procedures**], [**Count grouping item**]}

Fig. 4 An example of contextualized disease description frame.

ical terms to concepts described by CG based scheme permitted us to disambiguate and to improve explicitness and consistence of existing terms, to insert them at their right place within a hierarchy. This approach enables the integration of subsets of relevant medical terms coming from various terminologies in a structured poly-hierarchical semantic network devoted to support the domain ontology, in a manner close to the approach in the MAOUSSC project [31]. The use of precise rules to assess hyponymy permits to set up the hierarchy of terms on sound ontological foundations. Especially, we found that the systematic use of a discriminating associated disease to specify the chaining of primary, secondary and ternary diseases is very important to avoid errors in the hyper/hyponyms relations and thus helps a lot in the building of the ontology. The compositional principles related to the knowledge representational model we used are close to the "foundational model" as proposed by [32,33]. We used CG formalism as underlying knowledge representational model. Description logic, also a subset of first order logic and an inheritor of early semantic networks, is not very far from the CG model [25–27]. Other formalisms have been proposed for Knowledge Representation such as Knowledge Interface Format (KIF) [34], Arden Syntax [35], Ontolingua [36] or Grail, the KRF use in the core of the GALEN project [37,38]. Among other formalisms, some are very close to CG such as ENV 12265 or ENV 12264 proposed by CEN/TC251/WG2, inspired by the GALEN experience [37,38].

Our results with RIBOSOME demonstrate how the organization of a semantic lexicon according to a double paradigmatic and syntagmatic axis semiotic network provides an efficient support to integrate and to share a large amount of knowledge. This semiotic network is simultaneously the support of the internal ontology of the domain (the paradigmatic axis) and the support of the declarative rules for compositionality of conceptual structures (the syntagmatic axis). Conceptual structures are inserted into the semiotic network that permits the re-use of the acquired lexical and semantic knowledge for next analysis. This re-use appeared as an interesting mechanism for the automated completion of the disease description frames. With nearly 200 analyzed terms, we progressively gained in robustness and generality. The use of Natural Language Processing tools to set up domain ontologies is certainly a promising direction [39,40]. Such approaches realize a re-engineering of existing terminologies and set up the structure of terminological server on experimental basis.

Interestingly, our work on terminology and semantic heterogeneity in the domain of organ failure and transplantation raised the need to define the contexts of use of medical terms: information production is a communication process with a locutor, a message, a context of production of this message, a recipient and a context of exploitation of the message. We propose a solution recording the context within the terminological server, completing the inheritable and sharable disease description frames with local contextual slots.

According to the typology proposed by [29,30], we are building a terminological system with two classical goals: standardization and communication. Because EfG maintains a national information system devoted to evaluation and epidemiology, a third goal must be pointed out: the exploitation of the data. An inappropriate terminology is indeed a

source for information bias. The tuning of the lists of available medical terms usable for coding patients' information to the user's context is a pertinent requirement, facilitating the acceptability of the IS and improving the quality of information records. It may also help to minimize information bias. In one hand, when a user completing a form chooses an item within a list of medical terms, it does not mean that the items he did not choose are absent in the patient. This phenomenon may result in poor specificity and sensitivity of the coding procedure in use, with high discrepancies between users in their practice of the coding system. On the other hand, a systematic checking of yes and no boxes is not feasible for a large amount of coding items. Thus, an intermediate solution is the use of what we referred as to user's interface grouping items. A grouping item validates or invalidates the presence of its hyponyms according to the hierarchy. The presence or the absence of such items is systematically checked. Such grouping items will dynamically drive the construction of user's interface screens.

The implementation of the EfG-IS terminological server is now achieved, providing ontology editing functionalities (i.e. tools for the creation of new concepts, new relations and description frames), terminology management tools and users management tools (cognitician, medical expert, end-user). Further work is thus to record the entire ontology of the domain. A prototype of its integration within the information production system permits to evaluate the interest of having a contextual coding system.

Acknowledgements

Our thanks to the clinical experts who helped us in the validation of the semantic analysis validation process: A. Bisson, G. Bobrie, K. Boudjema, P. Landais, M. Stern.

References

[1] C. Jacquelinet, D. Houssin, in: J.L. Touraine, et al. (Eds.), Principles and Practice of Cadaver Organ Allocation in France, Kluwer Academic Pub, 1998, pp. 23–28.
[2] B. Stengel, P. Landais, Data collection about the case management of end-stage renal insufficiency, Nephrologie 20 (1999) 29–40.
[3] A. Sheth, J. Larson, Federated systems for managing distributed, heterogeneous and autonomous databases, ACM Computing Surveys 22 (1990) 183–235.
[4] P. Landais, A. Simonet, D. Guillon, C. Jacquelinet, M. Ben Saïd, C. Mugnier, M. Simonet, Un système d'information multi-sources pour le réseau épidémiologie et information en nephrologie: le projet REIN, in: Informatique et Santé, vol. 13, Springer, France, 2002, pp. 131–138.
[5] IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries, IEEE, 1990.
[6] A. Rossi Mori, F. Consorti, Exploiting the terminological approach from CEN/TC251 and GALEN to support semantic interoperability of healthcare record systems, Int. J. Med. Inf. 48 (1998) 111–124.
[7] W.E. Hammond, The role of standards in creating a health information infrastructure, Int. J. Biomed. Comput. 34 (1994) 29–44.
[8] J. Dudeck, Aspects of implementing and harmonising healthcare communication standards, Int. J. Med. Inf. 48 (1998) 163–171.
[9] G.W. Beeler, HL7 version 3—an object-oriented methodology for collaborative standards development, Int. J. Med. Inf. 48 (1998) 151–161.
[10] C.J. Mc Donald, et al., What is done, what is needed and what is realistic to expect from medical informatics standards, Int. J. Med. Inf. 48 (1998) 5–12.
[11] N.P. Mallick, E. Jones, N. Selwood, The European (European Dialysis and Transplantation Association-European Renal Association) Registry, Am. J. Kidney Dis. 25 (1) (1995) 176–187.
[12] G.R. Bramer, International statistical classification of diseases and related health problems. Tenth revision, World Health Stat. Q. 41 (1) (1988) 32–36.
[13] K. Innes, J. Hooper, M. Bramley, P. DahDah, Creation of a clinical classification. International statistical classification of diseases and related health problems: 10th revision, Australian modification (ICD-10-AM), Health Inf. Manage. 27 (1) (1997) 31–38.
[14] R.F. Averill, R.L. Mullin, B.A. Steinbeck, N.I. Goldfield, T.M. Grant, Development of the ICD-10 procedure coding system (ICD-10-PCS), Top. Health Inf. Manage. Feb. 21 (3) (2001) 54–88.
[15] K. Spackman, K.E. Campbell, Compositional concept representation using SNOMED: toward further convergence of clinical terminologies, Proc. AMIA Symp. (1998) 740–744.
[16] R.H. Dolin, K. Spackman, A. Abilla, C. Correia, B. Goldberg, D. Konicek, J. Lukoff, C.B. Lundberg, The SNOMED RT procedure model, Proc. AMIA Symp. (2001) 139–143.
[17] A.T. Mac Cray, The UMLS semantic network, in: Kingsland (Ed.), Proceedings of Symposium on Computer Applications in Medical Care, IEEE Computer Society Press, Washington, DC, 1989, pp. 503–507.
[18] P. Srinivasan, Exploring the UMLS: a rough sets based theoretical framework, Proc. AMIA Symp. (1999) 156–160.
[19] J.F. Sowa, Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, London, 1984.
[20] J.F. Sowa, Conceptual Graph Summary, in: Nagle, Nagle, Gerholz, Eklund (Eds.), Conceptual Structures: Current Research and Practice, Ellis Horwood Workshops, 1992.
[21] A.M. Rassinoux, R.H. Baud, J.R. Scherrer, Conceptual graph model extension for knowledge representation of medical texts, in: K.C. Lun, P. Degoulet, T.E. Piemme, O. Reinhoff (Eds.), MEDINFO 92: Proceeding of 7th World Congress on Medical Informatics, North-Holland Publ Comp, Amsterdam, 1992, pp. 1368–1374.
[22] K. Campbell, A. Das, M.A. Musen, A logical foundation for the representation of clinical data, J. Am. Med. Informatics Assoc. (1) (1994) 218–232.
[23] F. Volot, M. Joubert, M. Fieschi, Review of biomedical knowledge and data representation with conceptual graphs, Methods Inf. Med. 37 (1998) 86–96.
[24] A.-M. Rassinoux, R. Baud, C. Lovis, J. Wagner, J.R. Scherrer, Tuning-up conceptual graph representation for

multilingual natural language processing in medicine, in: M.-L. Mugnier, M. Chein (Eds.), Conceptual Structures: Theory, Tools and Applications. Lecture Notes in Artificial Intelligence, vol. 1453, Springer, Berlin, Heidelberg, New York, 2000, pp. 390–397.

[25] F. Volot, P. Zweigenbaum, B. Bachimont, M. Ben Saïd, J. Bouaud, M. Fieschi, J.F. Boisvieux, Structuration of medical knowledge. Using UMLS in the conceptual graph formalism, Proc. Annu. Symp. Comput. Med. Care (1993) 710–714.

[26] U. Hahn, M. Romacker, S. Schulz, How knowledge drives understanding—matching medical ontologies with the needs of medical language processing, Artif. Intell. Med. 15 (1) (1999) 25–51.

[27] J.F. Sowa, Knowledge Representation. Logical, Philosophical and Computational Foundations, Brooks/Cole (Eds.), Pacific Grove, 1999.

[28] C. Jacquelinet, A. Burgun, Building the ontological foundations of a terminology from natural language to conceptual graphs with RIBOSOME, a knowledge extraction tool. ICCS proceedings, in: G. Stumme (Ed.), Working with Conceptual Structures, Shaker Verlag, Aachen, 2000.

[29] N.F. De Keizer, A. Abu-Anna, J.H.M. Zwetsloot-Shonk, Understanding terminological systems I: terminology and typology, Methods Inf. Med. 39 (2000) 16–21.

[30] N.F. De Keizer, A. Abu-Anna, Understanding terminological systems II: experience with conceptual and formal representation of structure, Methods Inf. Med. 39 (2000) 22–29.

[31] A. Burgun, O. Bodenreider, P. Denier, D. Delamarre, G. Botti, P. Oberlin, J.M. Leveque, M. Bremond, M. Fieschi, P. Le Beux, A collaborative approach to building a terminology for medical procedures using a Web-based application: from specifications to daily use, Medinfo 9 (Pt 1) (1998) 596–599.

[32] A. Rector, Compositional models of medical concepts: towards re-usable application-independent medical terminologies, in: P. Barahona, J.P. Christensen (Eds.), Knowledge and Decisions in Health Telematics, IOS Press, Amsterdam, 1994, pp. 109–114.

[33] K. Spackman, K.E. Campbell, Compositional concept representation using SNOMED: toward further convergence of clinical terminologies, Proc. AMIA Symp. (1998) 740–744.

[34] M.R. Genesereth, R.E. Fikes, Knowledge Interchange Format, Version 3.0. Reference Manual, Computer Science Department, Stanford University, 1992.

[35] G. Hripcsak, P.D. Clayton, T.A. Pryor, P. Haug, O.B. Wigertz, J. Van der Lei, The Arden Syntax for Medical Logic Modules. Proceedings of the 14th Annual SCAMC, Washington, DC, 1990, pp. 200–204.

[36] T. Gruber, A translation approach to portable ontology specifications, Knowledge Acquisition 5 (1993) 199–220.

[37] A.L. Rector, A.J. Glowinski, W.A. Nowlan, A. Rossi-Mori, Medical-concept models and medical records: an approach based on GALEN and PEN&PAD, J. Am. Med. Informatics Assoc. 2 (1994) 19–35.

[38] A.L. Rector, A. Rossi-Mori, F. Consorti, P. Zanstra, Practical development of re-usable terminologies: GALEN-in-USE and the GALEN organisation, Int. J. Med. Inf. (48) (1998) 71–84.

[39] S. Le Moigno, J. Charlet, D. Bourigault, P. Degoulet, M. Jaulent, Terminology extraction from text to build an ontology in surgical intensive care, Proc. AMIA Symp. (2002) 430–434.

[40] B. Trombert-Paviot, J.M. Rodrigues, J.E. Rogers, R. Baud, E. van der Haring, A.M. Rassinoux, V. Abrial, L. Clavel, H. Idir, GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures, Int. J. Med. Inf. 58-59 (2000) 71–85.