Automatic Ontology Population extracted from SAM Healthcare Texts in Portuguese

David Mendes; Irene Pimenta Rodrigues

Departamento de Informática da Universidade de Évora

{dmendes;ipr}@uevora.pt

Abstract. We describe a proposal of the steps required to automatically extract the information about healthcare providing activities from an actual EHR[1] at use in a Portuguese Region (Portalegre) to populate an Ontology. We present the steps to manually and further automatically populate using a suggested Software Architecture and the appropriate Natural Language Processing techniques for Portuguese Clinical jargon.

1 Introduction

We will present in this paper our proposal on how to populate a clinical practice ontology, CPR[2], automatically from texts that can be obtained from the SAM [3] software system at use in ULSNA [4].

1.1 Motivation

The Semantic Web tools and techniques have come of age to be able to use an Ontology about a specific scientific and/or professional domain as knowledge representation scaffolding enough to be able to reason automatically and semantically inter-operate in that domain. The enormous amount of data residing in EHR renders the possibility of manually curating an ontology from that wealthy resource virtually impossible. We think that we dominate enough scientific knowledge about Natural Language Processing in general, and Portuguese in particular, to try applying it to the Healthcare providing domain in order to seriously contribute to the enhancement of the quality and cost effectiveness of that important activity.

1.2 Previous work

After researching the State-of-the-Art in knowledge acquisition from text in the Biomedical domain we presented some papers to peer-reviewed internacional

[1] Electronic Health Record

[2] Computer-based Patient Record Ontology

[3] Sistema de Apoio ao Médico - Doctors Support System

[4] Unidade Local de Saúde do Norte Alentejano - North Alentejo Local Health Unit

conferences about ontology enrichment in the heathcare domain. Several agreements with local heath authorities and healthcare providers have been signed in order to be able to develop work with reasonable corpora of data to demonstrate the applicability of the work of this research in real world situations.

1.3 Work in progress

We have colected some dozens of texts in PDF format that we use in this report both manually and automatically, as seen ahead, to enrich the CPR ontology. We are in the process of demonstrating the possibility of information extraction from free-text clinical episode reports to populate the ontology in an automated manner.

2 Experiences in the field

The ULSNA, E.P.E. [5] has as its principal object the provision of primary and secondary health care, rehabilitation, palliative and integrated continued care to the population. In particular to the beneficiaries of the national health service and the beneficiaries of the health subsystems, or with external entities with which it contractualized the provision of health care and to all citizens in General. Also articulate with public health activities and the means necessary to exercise the powers of the health authority in the geographic area affected by it. ULSNA, also has as it object to develop research, teaching and training activities. ULSNA is a healthcare providing regional system that includes 2 hospitals (José Maria Grande in Portalegre and Santa Luzia in Elvas) and the primary care centers in all the district counties: Alter do Chão, Arronches, Avis, Campo Maior, Castelo de Vide, Crato, Elvas, Fronteira, Gavião, Marvão, Monforte, Nisa, Ponte de Sôr, Portalegre e Sousel. Universidade de Évora signed an agreement with ULSNA that enabled the usage of de-identified (according to safe-harbor principles) clinical data from the SAM system in use both in the Primary Healthcare units and in the Hospitals. Using the clinical data that was available for us we populate the ontology for automatic reasoning capabilities over the suggested OWL2 ontology.

2.1 Acquisition points from SAM SOAP to CPR

In any EHR the number of direct sources for ontology instance retrieval is enormous. The number of registered clinical episodes from which we can populate our ontology induces any ontology engineer in a very hard problem to solve just trying to figure out the granularity of the ontology to be able to represent a realistic view from where valuable information can be extracted. When trying to apply the principles of well defined formal ontologies depicted in [5] and trying to avoid the errors mentioned in [1] we decided to get our hands wet with a

[5] www.ulsna.min-saude.pt

simple approach to the representation of disease and diagnostic as illustrated in [4].

In the SAM system there lies a clear support for text divided by 4 pre-defined subsections curiously acronymed SOAP after Subjective, Objective, Analysis and Plan. For any particular encounter (actually for any Clinical Episode) the text for any of these may be collected in the form of text suitable for processing into the Ontology and the appropriate suggested points in the CPR timeframe are the following:

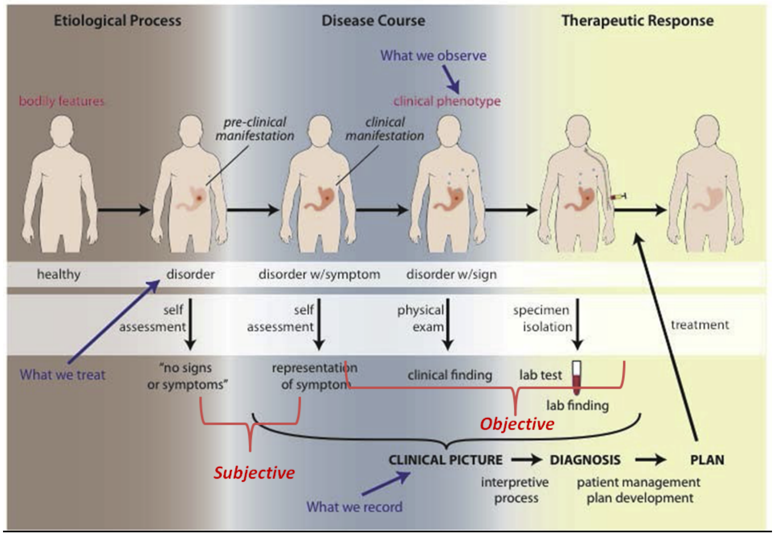


Fig. 1. SOAP Points Insertion

Where Subjective notes are those kind of signs normally expressed by the patient as well as soft validatable symptoms. Objective findings are all kinds of observations and results from quantifiable exams. Processing and populating the Ontology with the Analysis record labeled as Diagnosis the "Clinical Picture" is completed and is only lacking the Plan being instanced to render the therapeutic response associated with this encounter.

3 Automated acquisition from Clinical Episodes Text

3.1 The generic situation

As reviewed by the authors in Mendes and Rodrigues [2] the state-of-the-Art for acquisition from Clinical Text has enjoyed strong developments in recent years. In the mentioned paper we presented a proposal for automated acquisition from

HL7 messaging but here we are delving into the more generic possibility of extracting from free text present in most interfaces used by clinicians. Going from clinical episodes free text that is usually presented in a human friendly format to one adequate for computer processing involves a fair amount of text processing to handle situations like:

- Reports aggregate information from different clinical episodes that are not uniquely identified or not even individually dated
- The clinician is only identified by his/her name if any identification is made at all
- The information conveyed in free text is intended only to be understandable by fellow practitioners or even by the clinician himself making use of pragmatic jargon normally plagued with acronyms and nicknames abundant in their specific community
- Text is profoundly intermixed with decorative elements for better legibility, normally in PDF or HTML files
- The clinicians natural language is other than English without concepts defined in the foundational thesaurus like SNOMED CT or FMA for instance that don't exist in that particular language
- The time spanning of the processes depicted in natural language are difficult to represent formally

3.2 The adequate annotation workflow

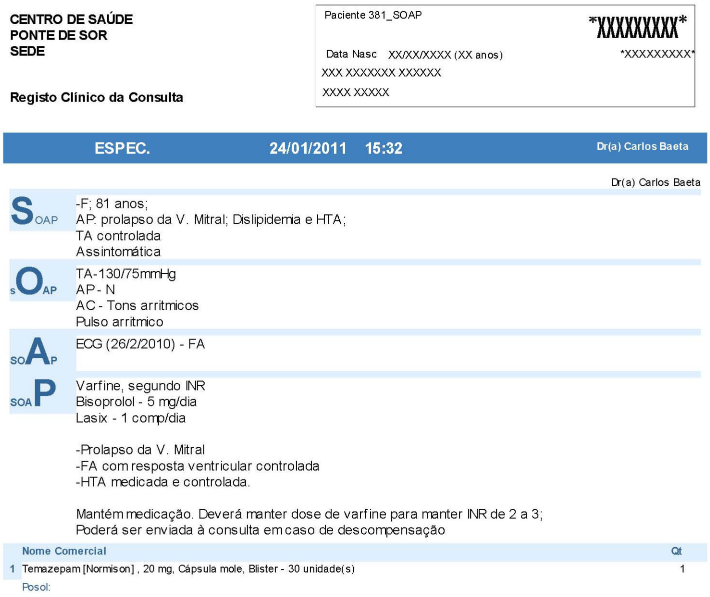
A set of sequential steps must be used to go from the pure text to the extracted CPR instance. Initially these tasks are done manually but after the initial proof of concept and tool customization they are automated. Those steps workflow can be configured declaratively using the software architecture shown in section 4. There are steps involved that consist of:

- PDF to raw text or to structured (XML) converting for adequate documents cleansing. For instance the graphical presentation of Vital Signs that are originally rendered in the respective report has to be deleted from the document for easier terms processing and the tables with values must be structured accordingly for the annotators to behave properly. Initially there is a proof of concept that involves manually cleaning the original reports
- Manual translation (that is indispensable for the translator tutoring as shown in 3.3) with the precise clinicians validation of their jargon adequately translated into English,
- Annotation using the Web interface of any of the services that we introduce in 3.5, either manually the interactive interfaces or automatically the Web Services available
- Filtering the concepts from the annotated text to insert in CPR instances

Given the array of available Web Services that can semantically annotate biomedical concepts in English in 3.5, we choosed to use an evolutionary approach for use of the BioPortal annotator [3]. We mean by evolutionary approach the fact that we first use the annotator after manual pre-processing and then a more automatic workflow.

3.3 The Multilingual problem

We can take advantage of the fact that we have to translate from jargon to English to customize the Google translator toolkit[6] with our own Translation Memories and Glossaries. Let us introduce some demonstrative examples taken from a sample document gently provided by Dr. Carlos Baeta and properly de-identified:



CENTRO DE SAÚDE
PONTE DE SOR
SEDE

Registo Clínico da Consulta

Paciente 381_SOAP

XXXXXXXXX

Data Nasc XX/XX/XXXX (XX anos)

XXXXXXXXX

XXX XXXXXXX XXXXXX

XXXX XXXXX

ESPEC. **24/01/2011 15:32** **Dr(a) Carlos Baeta**

Dr(a) Carlos Baeta

S OAP

-F; 81 anos;
AP: prolapso da V. Mitral; Dislipidemia e HTA;
TA controlada
Assintomática

S O AP

TA-130/75mmHg
AP - N
AC - Tons arritmicos
Pulso arritmico

SO A P

ECG (26/2/2010) - FA

SOA P

Varfine, segundo INR
Bisoprolol - 5 mg/dia
Lasix - 1 comp/dia

-Prolapso da V. Mitral
-FA com resposta ventricular controlada
-HTA medicada e controlada.

Mantém medicação. Deverá manter dose de varfine para manter INR de 2 a 3;
Poderá ser enviada à consulta em caso de descompensação

	Nome Comercial	Qt
1	Temazepam [Normison] , 20 mg, Cápsula mole, Blister - 30 unidade(s)	1

Posol:

Fig. 2. SOAP 381

We will, in the process of using the Google toolkit, create Translation Memories with the identified personal acronyms like:

- AP (Antecedentes Pessoais) into Personal History
- HTA (Hiper Tensão Arterial) into High Blood Pressure
- FA (Fibrilhação Auricular) into Atrial Fibrilation
- V. Mitral (Válvula Mitral) into Mitral Valve

Some which are acronyms that can be given the suitable translated concept like:

- ECG (Electro Cardio Grama) into Electro Cardio Gram

or those that are even English acronyms:

[6] https://translate.google.com/toolkit

- INR (International Normalized Ratio) into International Normalized Ratio

and some which are not really necessary because the conveyed information is irrespective of what language is in like:

- CENTRO DE SAÚDE into HEALTH CENTER
- SEDE into MAIN OFFICE

Included in this sample are notorious some more complex problems that are not related to the translation itself but with some other problems like the time spanning of concepts like “1 comp/dia” which is adequately translated to “1 tablet per day” using the defined Translation Memory but has to be posteriorly well defined as time delimited occurring process this kind of problems and the suggested solutions are itemized ahead in 3.4.

3.4 SAM Corpus

In our particular case we face a shortage of structure of the reports extracted from SAM in order to be able to fill the suitable instances in CPR in a more systematic way. As an example we illustrate below that no complete demographic information is available in any of the used reports or that the problems enumerated in the Problem List don't have any kind of severity or progress information associated. Reports that are produced by the SAM system and have interest for our work are:

- List of Medical Problems

This report lists dated references to medical problems that were found to be of reporting interest at any particular moment in the patients live. It provides a history line of clinical problems without any relation to any procedures taken whatsoever. One of the interesting contributions of automatic acquiring this information into our ontology is the possibility to generate a problem oriented timeline into which all the further discovered clinical procedures can be related to. It lacks, however, any kind of data referring intensity or gravity of the problem or its evolution along the time frame from when it was declared active until the problem dismissal. So far, as what is of interest to our Ontological Realism approach, it can be seen as a possibility of reasoning by fact inferring and it renders then our proposal somewhat more interesting because that possibility can be easily demonstrated. The pre-processing step of this type of documents has to take into appropriate care the fact that concepts extracted can be found in a “feeder ontology” like MeSHPOR and thus have an ID from a standard vocabulary associated or not. From problem we can then extract possibilities that will allow us to identify if there is an associated etiologic agent, a pathological-disposition or a mere sign recording for instance and then give the possibility to refine the medical problem instance with further information that shall always agree to this pre-defined structure:

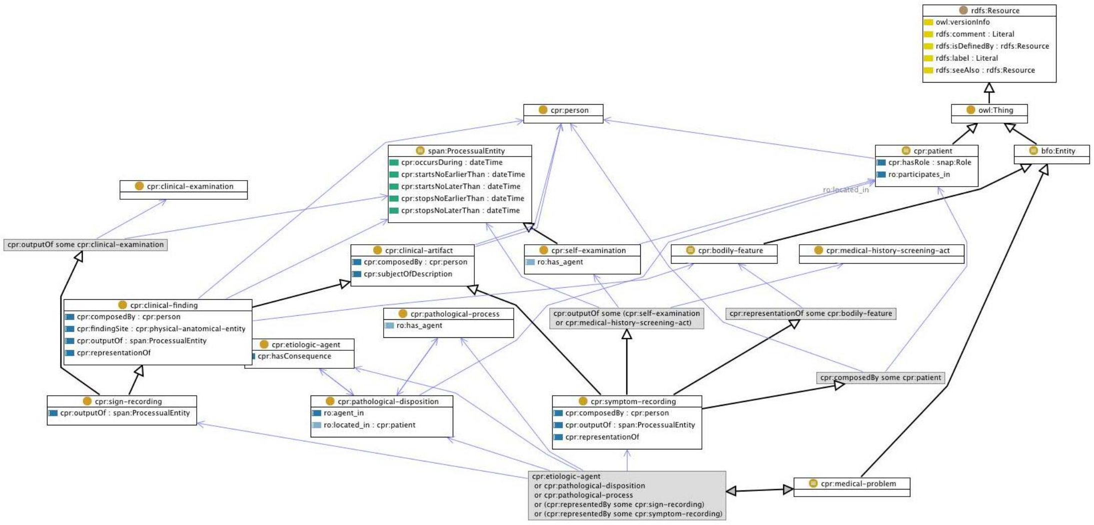

rdfs:Resource
owl:versionInfo
rdfs:comment : Literal
rdfs:isDefinedBy : rdfs:Resource
rdfs:label : Literal
rdfs:seeAlso : rdfs:Resource
owl:Thing
cpr:person
span:ProcessualEntity
cpr:occursDuring : dateTime
cpr:startsNoEarlierThan : dateTime
cpr:startsNoLaterThan : dateTime
cpr:stopsNoEarlierThan : dateTime
cpr:stopsNoLaterThan : dateTime
cpr:patient
cpr:hasRole : snap:Role
ro:participates_in
bfo:Entity
ro:located_in
cpr:clinical-examination
cpr:outputOf some cpr:clinical-examination
cpr:clinical-artifact
cpr:composedBy : cpr:person
cpr:subjectOfDescription
cpr:self-examination
ro:has_agent
cpr:bodily-feature
cpr:medical-history-screening-act
cpr:clinical-finding
cpr:composedBy : cpr:person
cpr:findingSite : cpr:physical-anatomical-entity
cpr:outputOf : span:ProcessualEntity
cpr:representationOf
cpr:pathological-process
ro:has_agent
cpr:outputOf some (cpr:self-examination or cpr:medical-history-screening-act)
cpr:representationOf some cpr:bodily-feature
cpr:etiologic-agent
cpr:hasConsequence
cpr:composedBy some cpr:patient
cpr:sign-recording
cpr:outputOf : span:ProcessualEntity
cpr:pathological-disposition
ro:agent_in
ro:located_in : cpr:patient
cpr:symptom-recording
cpr:composedBy : cpr:person
cpr:outputOf : span:ProcessualEntity
cpr:representationOf
cpr:etiologic-agent or cpr:pathological-disposition or cpr:pathological-process or (cpr:representedBy some cpr:sign-recording) or (cpr:representedBy some cpr:symptom-recording)
cpr:medical-problem

Fig. 3. CPR Medical Problems

This medical problems are, of course, an cpr:hipothesizedProblem of a cpr:clinical-diagnosis as we can see in the diagnoses view of the CPR ontology:

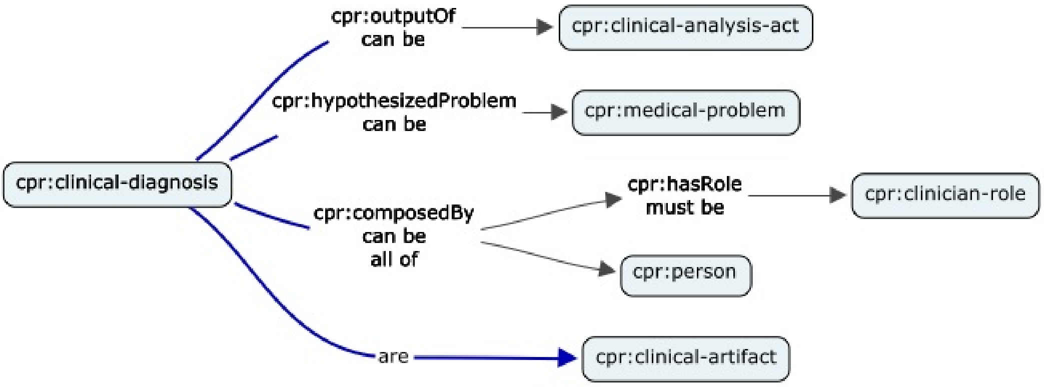


Fig. 4. CPR Diagnoses view

– SOAP Report

This report is the most informative of all available as it depicts a clinical encounter in a semi-structured way. As seen previously in the figure in section 3.3 we find sections that can be associated with

Symptoms, the subjective section S where we extract directly to cpr:symptom-recording.

Signs, the objective section O that are cpr:sign-recording that we take as generator for cpr:clinical-findings.

Actions, the analysis section A which are the cpr:clinical-investigation-act whose outputs can be cpr:clinical-artifact to investigate things that can be cpr:isConsequenceOf any of cpr:physiological-process or cpr:pathological-process

and finally

Plan, the plan section P where the therapeutic acts can be extracted with all the timing, posology and prescriptions registered in a particular clinical encounter.

– Different Areas Exams

This report has a summary of the different diagnostic complementary exams that a certain patient has been subjected to. It is divided into Laboratory Analysis, Pathological Anatomy, Common Exams and Imagiology. All these cpr:clinical-analysis-act contribute, or not, to a cpr:clinical-diagnosis that is a cpr:hypothesizedProblem of a cpr:medical-problem. This can be the workhorse of the automatic acquisition because the reports that are based in free-text can have origin from the different EHR modules and validated in advance rendering a high degree of certainty and even the possibility of pre-encoding according to some controlled vocabulary like ICD-9 or ulterior for example.

– Vital Signs Summary

The Vital Signs Report has to have the biggest amount of text cleansing of all because all the graphics have to be taken out and the tables have to be formatted as such for the annotators to understand them as tables and do the appropriate treatment that is take each line at once and create the specific instance in cpr:clinical-finding.

3.5 Services to annotate clinical concepts in free text

Apart from being able to provide “our own” Web Services for various tasks given the availability of downloading several types of terminologies like MeSH or SNOMED CT CORE and generally the UMLS Metathesaurus, currently there are a miriad of WS at reach that can be configured to be connected to our CP-ESB as service providers. Among those we think that are worth mentioned here the BIOPortal [7], OntoCAT[8] and UTS[9].

All these provide Web Services that offer specific tasks for Biomedicine terminology. Carefully chosen endpoints provide features that range from simple term lookups to complete semantic concept acquisition. Normally all these offerings are available at no cost upon registering and access granting.

[7] http://bioportal.bioontology.org

[8] http://www.ontocat.org

[9] https://uts.nlm.nih.gov/home.html

4 Software Architecture

To have an extensible architecture able to build upon and capable of entailing the current available tools we chose to use Java based tools. Namely building upon an Eclipse based modularization platform called OSGi[10]. The OSGi technology is a set of specifications that define a dynamic component system for Java. These specifications enable a development model where applications are (dynamically) composed of many different (reusable) components. The OSGi specifications enable components to hide their implementations from other components while communicating through services, which are objects that are specifically shared between components. This surprisingly simple model has far reaching effects for almost any aspect of the software development process.

Though components have been on the horizon for a long time, so far they failed to make good on their promises. OSGi is the first technology that actually succeeded with a component system that is solving many real problems in software development. Adopters of OSGi technology see significantly reduced complexity in almost all aspects of development. Code is easier to write and test, reuse is increased, build systems become significantly simpler, deployment is more manageable, bugs are detected early, and the runtime provides an enormous insight into what is running.

The OSGi technology was aimed to create a collaborative software environment. Here an application emerges from putting together different reusable components that had no a-priory knowledge of each other. The goal is to allow the functions to be added without requiring that the developers have intricate knowledge of each other and let the components be added independently.

One of the major architectural components that fosters the decoupling of the different components is a common rail where messaging can flow using a subscription model that enables the communication to be detached from any two particular services but instead be available on-request by one and served by another in a loosely coupled way. ESB[11] is a modular and component based architectural component. It assumes that services are generally autonomous and availability of a service at a certain moment of time cannot be guaranteed. Therefore messages need to be routed consequently through the message bus for buffering (message queuing to allow inspection and enhancement of content as well as filtering, correction and rerouting of message flow. In an enterprise architecture that makes use of an ESB, an application will communicate via the bus, which acts as the single message turntable between applications. That approach reduces the number of point-to-point connections between communicating applications. This, in turn, makes impact analysis for major software changes simpler and more straightforward. By reducing the number of points-of-contact from and to a particular application, it is easier to monitor for failure and misbehavior in highly complex systems and allows easier changing of components.

[10] www.osgi.org

[11] Enterprise Service Bus

It is an essential design concept of an ESB that every client directs all its requests through the ESB instead of passing it directly to a potential server. This indirection allows the ESB to monitor and log the traffic. The ESB can then intervene in message exchange and overwrite standard rules for service execution. The case of an intervention here is in the ability to filter and redirect invocations to the appropriate NLP task processors depending on the source being labeled with the kind of load it carries. For example, if an invocation carries the language labeled as PT it will invoke the MeSH concept translator before invoking the rest of the NLP processing pipeline. Another example may be workflow maintained in the ESB itself of the invocation of the CPO refiner before the CPO populator. The pipeline is maintained by configuration of the ESB and not hard-coded in any way. Buffering and delaying message in a staging area and automatically deliver it when the receiver is ready, monitor messages and services to be well-behaved, enforce compliance with dynamic processing and security policies, marshal service execution based on dynamic rules, prioritize, delay, and reschedule message delivery and service execution, write logs and raise exception alerts all are examples of the ESB workhorse functionality. Notably the REST kind of software architecture, that has in the World Wide Web its most prominent example, is suggested in our proposal as the correct way of implementing a SOA [12] that serves as the communication underlying structure of our system. REST interfaces are available for consuming for the generality of our needs as shown ahead. In the next figure we present the alignment of the process invocation to fill the points suggested in [4] for disease and diagnosis representation in our CPR ontology.

CP-ESB The advantages of using a Software component as important as the ESB in the current SOA world of application composition is beyond the scope of this work. An important source of information to get up-to-date might be the Wikipedia page: http://en.wikipedia.org/wiki/Enterprise_service_bus. In our case the applicability of technology to the Clinical Practice environment can be seen as:

[12] Service Oriented Architecture

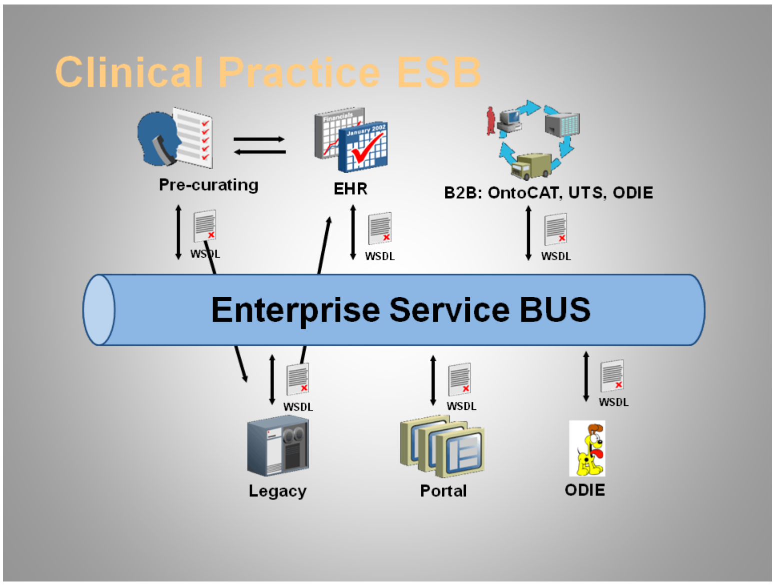


Fig. 5. Clinical Practice - ESB

Where the articulation of the providers and the Ontologies are well defined and are not handled point-to-point but always through the ESB routing and intercepting capabilities. All the core features are already implemented to enable plug-and-play ability for interchanged modules as stated above.

Connecting the dots Building over the suggested infrastructure the systems are rather composed as opposed to monolithicaly built and so manifest high capabilities of plug-and-play configuration allowing for interchangeable providers (as Web Services), Reference Ontologies (Feeders), and target ontologies. Having the foundations available with the right weapons provided one has to take a practical approach to the development of a target system using, in our case, the Java best-practices for pragmatic development that include a number of Patterns as in JEE [13] compiled in http://java.sun.com/blueprints/corej2eepatterns/Patterns/ or the pragmatic approaches developed in such successful projects as OSGi or Spring[14].

The flowchart that depicts graphically the acquisition from the source texts in Portuguese to the creation of the appropriate CPR instance is:

[13] Java Enterprise Edition

[14] http://www.springsource.com/

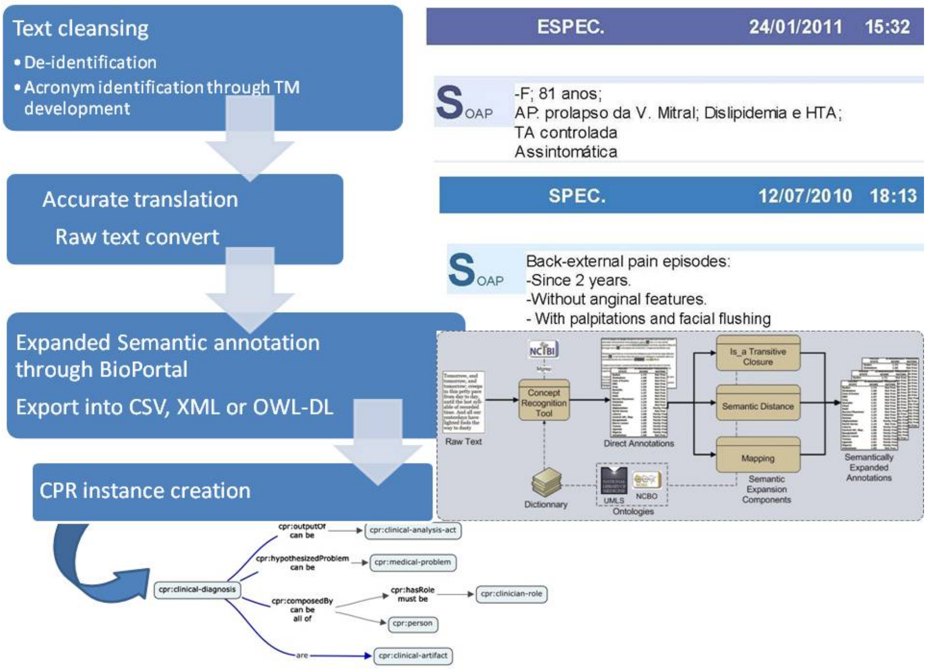


Fig. 6. Acquisition Flowchart

ODIE stands for "Ontology Development and Information Extraction". It is a software toolkit that uses ontologies to perform information extraction tasks from clinical documents and uses clinical documents to enhance existing ontologies. Both ODIE and BioPortal code document sets with ontologies or enrich existing ontologies with new concepts from the document set. They contain algorithms for Named Entity Recognition, Co-reference resolution, concept discovery, discourse reasoning and attribute value extraction. They allow development of reusable software leveraging existing NCBO tools and compatible with NCBO architecture. A downloadable version gives the possibility of developing local extensions to the algorithms provided in the base offering allowing, for instance, targeting different languages in the NLP tasks. The WS provided by BioPortal or OntoCAT can be locally extended and refined for all the sources are provided as one of the projects deliverables.

5 Conclusion

We presented a humble contribution to demonstrate the articulation needed of different software tools and medical knowledge to be able to fill a Clinical Practice Ontology with instances collected automatically from reports taken from a specific local EHR system.

Bibliography

[1] Ceusters, W., Smith, B., Kumar, A., Dhaen, C., 2004. Mistakes in medical ontologies: where do they come from and how can they be detected? Stud Health Technol Inform.

[2] Mendes, D., Rodrigues, I., 2011. A Semantic Web pragmatic approach to develop Clinical ontologies, and thus Semantic Interoperability, based in HL7 v2.xml messaging. In: HCist 2011 - Proceedings of the International Workshop on Health and Social Care Information Systems and Technologies. Springer-Verlag - book of the CCIS series (Communications in Computer and Information Science).

[3] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., Musen, M. a., Jul. 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research 37 (Web Server issue), W170–3.
URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2703982&tool=pmcentrez&rendertype=abstract

[4] Scheuermann, R. H., Ceusters, W., Smith, B., 2009. Toward an Ontological Treatment of Disease and Diagnosis. In: 2009 AMIA Summit on Translational Bioinformatics. San Francisco, CA, pp. 116–120.

[5] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., Lewis, S., Nov. 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology 25 (11), 1251–5.
URL http://www.ncbi.nlm.nih.gov/pubmed/17989687