Can an Ad-hoc ontology Beat a Medical Search Engine? The Chronious Search Engine case.

Piero Giacomelli
Tesan S.p.A., Italy
email:giacomelli@tesan.it

Giulia Munaro
Tesan S.p.A., Italy
email:munaro@tesan.it

Roberto Rosso
Tesan S.p.A., Italy
email:rosso@tesan.it

Abstract—**Chronious is an Open, Ubiquitous and Adaptive Chronic Disease Management Platform for Chronic Obstructive Pulmonary Disease(COPD) Chronic Kidney Disease (CKD) and Renal Insufficiency. It consists of several modules: an ontology based literature search engine, a rule based decision support system, remote sensors interacting with lifestyle interfaces (PDA, monitor touch-screen) and a machine learning module. All these modules interact each other to allow the monitoring of two types of chronic diseases and to help clinician in taking decision for care purpose. This paper illustrates how the ontology search engine was created and fed and how some comparative test indicated that the ontology based approach give better results, on some estimation parameters, than the main reference web search engine.**

Keywords- Telemedicine; chronic disease management; ontology search engine.

I. INTRODUCTION

Scientific advances over the past 150 years, particularly in the medical field, have allowed the extension of life expectancy in western countries and this trend seems to increase in future years. Conservative estimates suggest that by 2030 in EU countries the proportion of people over 60 years regard the entire population will be around 50%; this means that we will see a gradual increase in the number of those subjects with chronic diseases (i.e., diseases not involving healing), that will therefore increase the cost and effort over health care facilities [1].

Chronic diseases are slowing but constantly replacing malnutrition and infection as primary causes of mortality in the population [2]. The World Health Organization (WHO) has recently emphasized that chronic diseases are a global priority [3].It was calculated that, if governments are able to put in place public health policies that produce a 2% yearly reduction in mortality rates for chronic diseases, 36 million deaths would be prevented worldwide between 2005 and 2015 [4].

Chronic diseases are difficult to treat and, apart from deaths, have collateral social impact that are becoming an economic emergency both in western and developing countries. As the number of patient with chronic diseases is rising there will be an increasing cost for hospitalization structure both public and private. Considering some specific diseases like Chronic Kidney Disease (CKD), sometimes there is,

during the medical treatment, a non-return point from where the hospitalization is continuous as for dialysed people. The traditional approach consisting in periodic check-ups and periodic lab exams seems a model that won't be sustainable as the population gets older and the total number of patients with chronic diseases rises. At present the physician deals with an increasing number of chronic patients that are lowering the periodic check-ups and so reducing the ability to prevent, if not death, worsening in patient's quality of life.

In the latest years, we have seen a tremendous growth in IT infrastructure, both from the hardware and communication capacity. Nowadays a common mobile phone is much more powerful in terms of hardware and software capacity than the first calculating machine that allowed the man to land on the moon forty years ago. The continuous growth of the World Wide Web (WWW) and, linked to this, the continuous growth in bandwidth capacity for data transmission allows to have cheaper and more widely available bandwidth, for larger portions of the population.

As consequence of the exponential growth of hardware and software infrastructure it is possible to rethink the whole approach to the treatment of complex chronic diseases by limiting the hospitalization only to situations of severe worsening of patient condition. This was the original idea behind the EU funded Chronious project [5]: constructing a generic platform to monitor, in an unobtrusive way, patient with chronic disease in two goals [6]:

- Improve the patients quality of life, by reducing as much as possible the hospitalizations.
- Allow the clinician a continuous monitoring of the patients, both in standard and potential risk situations.

To gain this two goals, the Chronious platform has to integrate different technologies such as hardware and software modules that need to interact among themselves. This paper is focused on the ontology search engine module: we will illustrate what are the aims of this module and what are the main components. The storage system for the documents is an ontology, developed specifically for the CODP and CKD diseases. We will illustrate how this ontology was created and enriched, from medical literature

sources. Finally we will illustrate our tests conducted against the principal medical search engine and how the preliminary results seem to indicate that such approach can outperform the results of a web search engine. The paper is organized in the following sections. We will first describe the Ontology Chronious Search Module and what were the needs and how we solved them. Chapter three fully illustrate the whole process of document uploading and processing of the text to gain the enrichment of the ontology. Finally we will illustrate our tests results and suggest some future improvements.

II. CHRONIOUS SEARCH MODULE

Chronious is an hardware/software platform devoted to monitor in a remote way COPD and CKD patients. In Figure 1, a schema of the whole system is presented:
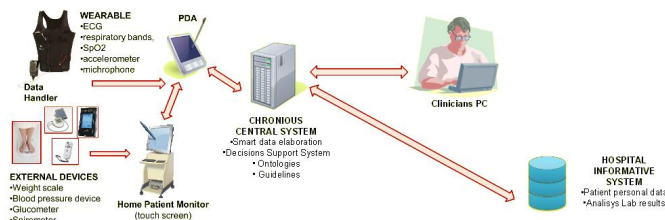

Figure 1.   Chronious modules

The chronic patient is equipped at home with the following devices:

- a Personal Digital Assistant (PDA) that contains machine learning algorithms, acting as a first alerting system and that is used to transmit data to a central system through GPRS technology.
- A Home Patient Monitor (HPM) that is used to allow patients to insert qualitative information on diet, activity and drugs intake.
- Bluetooth medical devices: weight scale, blood pressure intake device, glucometer, air quality monitoring tool. Coupled with these devices a sensorized t-shirt for vital parameters recording like cardiac and respiratory signals.

For the two pathologies, different sets of equipments are given to the patient according to the clinician. For example the glucometer is usually assigned to CKD patient affected by diabetes comorbidity .

The data collected are automatically transmitted via GPRS to a Central System that, using a web interface, and a rule-based Central Decision Support System (CDSS), allows clinician to monitor patient and to receive suggestions on how to act in case of a worsening trend or a potentially life risk situation.

The CDSS uses JENA [7] framework and a set of rules codified in OWL [8] format to display suggestions to the clinician. An example of such a rule is displayed in table II.

Table I
EXAMPLE OF A SUGGESTION GIVEN BY THE CDSS

patientID	1
AlertType	White
Date	01/02/2011
Description	body temperature up to 38
Suggested action	hospitalization of the patient
Guideline Text	text from literature

For CKD we used the KDOQI Guidelines [9], for COPD we create a set of rules based on clinician experience. The suggestions are portion of text extracted from literature reference or documents provided by the experts. Managing documents and information , and organising concepts, such as comorbidities, and relations between them is a central task, for having a correct outcome to the CDSS calling.

Searching literature reference, we find that latest information retrieval/storage systems studies, seem to indicate that ontology structure are a better way details concepts and the relations between them [10]. Ontologies have been used successfully on medical [11], genetic [12] and surgery [13] fields. The document repository for storing informations on the COPD and CKD diseases, was chooses as an ontology. This lead us to other issues: how to build the ontology and how to enrich it with new concepts and to validate it.

Agreeing with doctors , the main reference on the medical field is the PubMed[14] site. PubMed is a free database accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics. As of 1 July 2011, PubMed has over 21 million records going back to 1966, selectively to the year 1865, and very selectively to 1809. About 500,000 new records are added each year. As of 1 July 2011, 11.9 million articles are listed with their abstracts and 3.3 million articles are available full-text for free.   The problem was how to integrate all this huge among of information in the PubMed site with the Chronious Ontology. Apart from the integration the main problem was how to connect all sparse information about COPD and CKD pathologies in a way to produce a real value information to the clinician. Querying Pubmed for generic COPD (at 1th September 2011), we have as outcome 33319 documents.

We coupled the Chronious Documents repository with a Search Module for having an interface that clinician can use to fast access literature specific information on the two diseases treated by Chronious.

Using these intuitions, the whole Chronious Search Module of the Chronious System is composed of four main parts:

- Upload tool: a web interface that is able to upload files to the repository.
- The repository itself and the ontology used to underline concepts and relation extracted from the raw text.
- Enrichment tool: a tool that after the text information extraction is able to say which new concepts and relations should be accepted.

- A search tool. This is the final web interface for the clinician that using the concepts and terms provided in a free text search or more structured way is able to query the repository and give documents as feedback.

Being that the Chronious Search Tool was the entry point for the CDSS, the clinician feedbacks about the whole architecture underlined potential issues. The main issue, was that it is useless to use a personalized search engine while going via web to PubMed was so fast and so accessible. Another issue was that the clinician should upload manually the document into the repository. We will describe the processing of a single document to demonstrate how we solve these issues.

III. THE CHRONIOUS UPLOAD/PROCESSING SYSTEM

The Chronious ontology have been developed using the DOLCE ontology [15], however to enrich the relation between the nodes of the ontology graph, a functionality for transforming text into concept were needed. The upload functionality is based on a web interface that is able to upload a pdf file and associating it with some general information like author, journal, year, volume. Due to the fact that the more documents are uploaded the better the ontology can be enriched, we code a web spider that is able to use Pubmed site structure to download automatically.

Even if the final step of the process, the Enrichment Tool validation, needs a human approval of the concept chosen, this automated download tool helps the boring part of downloading a document from PubMed and uploading to the Chronious system.

The automated download tools is basically a web crawler that periodically use a query string over the Pubmed search functionality, do the HTTP request, parse the html pages and find html tag for having information about the pdf file to download. The code was developed using .NET framework 3.5 being that the CLR provide a library that is able to interact directly with an html page using the DOM (Document Object Model) system.

This allows easily to mimic nearly every interaction that the human user can have with the page like: entering text in the input fields and clicking on a link to simulate the "save as" functionality.

After the downloading finished successfully the information are stored on a RDBMS database so that only new papers are downloaded and parsed.

One of the problem faced with this approach is the copyright issues that affects the contents of the PubMed database. The greatest part of the articles indexed by PUBMed are protected by copyright, as they come from journals that have copyright agreements, so in most of the case the content of the document cannot be viewed for a user that have no subscription. The problem is not present if we use the upload tool is used inside an institution that have a subscription with PubMed based on the ip address.

However if the Automated Upload Tool is installed on a normal pc downloading document provide an infringement of the PubMed copyright.

To avoid this we decide to use a lower set of documents that are provided by Pubmed for free until the document is published on the journal, for having a first test on the system and to see if the concepts extracted were in line with the ontology we build. To prevent any possible copyright issue, we decide to show to the final user only the DOI of the document so in case the user would like to see the whole document he must open a browser window. This action will shift the copyright from the Chronious system to the final user of the search tools. At the end of the enrichment process the module removes the physical pdf. Once the document is uploaded into the repository a Natural Language Processing (NLP) is used to for Information Extraction (IE). For implementing the NLP algorithms the GATE [16] framework will be used. GATE is a leading infrastructure for developing and deploying software components, that process human language. Among others it provides a framework, based on JAVA, that implements the architecture and can be used to embed language processing capabilities in different applications.

GATE supports many document formats like: Plain Text, HTML, SGML, XML, RTF, Email, PDF and Microsoft Word.

Every text is splitted into sentences and words and and every concept extracted is then indexed and associated with a weight that evaluate the correspondence with the other concepts already present into the Chronious ontology.

Once this evaluation is done the concept extracted are presented to the human user using a web interface, where the new concepts are shown with their evaluation indexes see Figure 2.
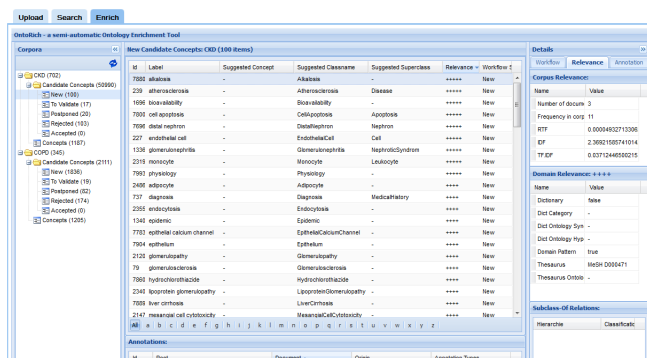

Figure 2.   Enrichment Tool

After the whole process finished, a web interface that able to query the ontology for terms evaluation in a structured or free text search form (see Figure 3) is the last interface. So a clinician is able to use it for fast finding literature reference. In addition, the search functionality, uses also the concepts

in the ontology so it is possible to have fast reply to query like "abnormal coughing toxicity".
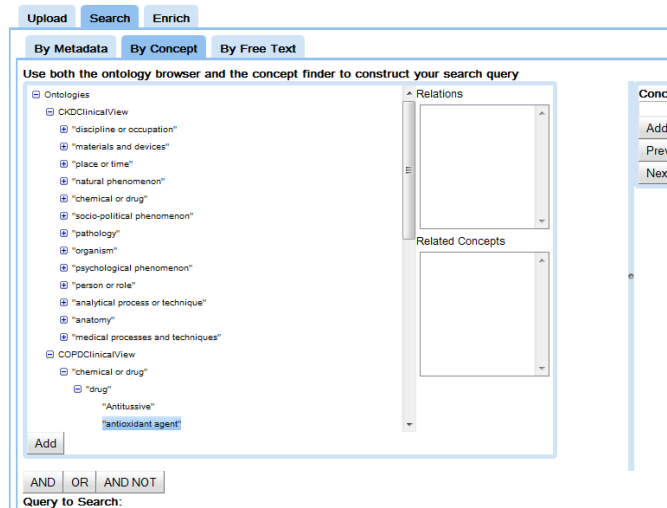

Figure 3.   Search Tool

The goodness of the query results should finally be evaluate. Next chapter will detail the criteria used.

IV. EVALUATION

As we said having another search functionality apart from PubMed raised some murmurs from clinicians, so we need to show that from the perspective of information retrieved, the Chronious Search Tool has some advantages over PubMed.

For doing this, various classifications and criteria for ontology evaluation have been considered in the literatures [17]. However, no standard evaluation criteria have been defined so far. Reference literature proposed [18] a "three dimensions" classification to evaluate ontology in three categories: Structural Dimension, Functional Dimension and Usability-Profiling Dimension.   Considering that Structural Dimension and Usability-Profiling Dimension where validated by ontology experts we focused on Functional Dimension with three different tests see Table II:

Table II
FEATURES TO BE TESTED IN CHRONIOUS ONTOLOGY TESTING

Test Identifier	Description
TF-B-1	Agreement Among Experts
TF-B-2	User-Satisfaction
TF-B-3	User-Satisfaction and Completeness of Search Result

The questionnaire interview approach can be applied in TF-B-1 to conclude the agreement among medical experts to evaluate the correctness of the developed Chronious ontologies. Test participants for this test were medical experts, especially in CKD and/or COPD area.

The black-box test TF-B-2 utilizes the same approach as TF-B-1, the questionnaire interview approach, to measure the overall satisfaction of the end-user. Because there are three different search options in the Search Module (Search by Metadata, Search by Concept, and Search by Free Text), the questionnaire should take all these options into account and compare their search result quality with each other. The people involved for this interview were the end-user of Chronious Healthcare Professional GUI, i.e., the medical experts.

In black-box test TF-B-3, Precision P is defined as the number of relevant documents retrieved by a search $g(r)$ divided by the total number of documents retrieved by that search N.

$$P = \frac{g(r)}{N}$$

while Recall R is defined as the number of relevant documents retrieved by a search $g(r)$ divided by the total number of existing relevant documents G (which should have been retrieved).

$$R = \frac{g(r)}{G}$$

Precision and Recall are two widely used statistical classifications, especially in information retrieval domain. Precision can be seen as a measure of exactness or fidelity, whereas Recall is a measure of completeness.

Usually, Precision and Recall scores are not discussed in isolation. Instead, either value for one measure are compared for a fixed level at the other measure (e.g., Precision at a Recall level of 0.75), or both are combined into a single measure, such as the F-measure [19], which is the weighted harmonic mean of Precision and Recall:

$$F = \frac{(\beta^2 + 1) \cdot P \cdot R}{(\beta^2 \cdot P) + R}$$

Whereby β is a value between 0 and 1 reflecting the weighting of Precision vs. Recall.

Before doing an evaluation we uploaded into the repository 1000 free access documents grabbed by the Automatic Upload Tool from Pubmed using two different query strings

- CKD treatment for CKD ontology part.
- COPD treatment limited to year 2008 for COPD ontology part.

Because the document repository of Chronious Search Module and of the PubMed Central system is artificially identical (constrained with some specific limitations), for same search query, the two systems should contain the same number of relevant documents to this query, i.e., G for that query in both systems should be equal. It follows that, no matter which value the variable G has, the comparison of F-measure between these two search systems will only be influenced by $g(r)$ and N of the search result. Hence, if the value of $g(r)$ and N are available for a search both with Chronious Conceptual Search option and with PubMed Central system, the F-measure of the search results with

both systems can be compared with each other with the help of their function diagrams (the search query used in both systems must be identical).

Some outputs of our tests can be seen in Table III

Table III
FEATURES TO BE TESTED IN CHRONIOUS ONTOLOGY TESTING

G	F-measure Chronious	F-measure Pubmed
500	0.724637681	0.595238095
1000	0.531914894	0.426136364
2000	0.347222222	0.27173913

The table shows that, the Chronious Conceptual Search option performs better than the PubMed Central system with the search query, no matter how many relevant documents existed in the repository to this query. With this approach, although the F-measure value of a search result cannot be estimated for every type of query, the search performance of the Chronious Search Module, however, can be compared with the search performance of the PubMed Central system, using the F-measure.

V. CONCLUSION AD FURTHER IMPROVEMENTS

Even if the test results seem promising, it is not possible to say that in general the ontology search approach can outperform a search functionality like PubMed one. However the good news, is that such kind of approach on storing information, is promising for developing future artificial intelligence systems for application in telemedicine. Linking concepts like symptoms to drug or caring procedures with relations underline by literature studies can greatly help clinician during their daily routines. It would be interesting to evaluate the Conceptual Search on the whole corpus of the Pubmed database even if this would be impossible due to copyright restriction and infrastructure storage possibilities. For sure PubMed will remain the main reference literature search engine, however it is our opinion that having structures information like the one in Chronious ontology, will for sure be an added value.

Another interesting task would be the evaluation of a complex free text query search over the two systems. Free text search remains the first approach for searching information. Reflection on how to benefit from ontology structured data in improving outcome for free text search seems is a research problem that require a deeper evaluation.

Last we point out that Chronious Enrichment Tool still need to have a human interaction. A still open research task, is the one of having some kinds of automation in inserting the new concepts and relations in the existing ontology, in way to have a sort of unsupervised concepts enrichment that mimics the rule extraction in transaction datasets. We leave these suggestions hoping that the reader will be interested to think about them.

REFERENCES

[1] C. Zoccali, A. Kramer, and K. Jager, "Chronic kidney disease and end-stage renal disease a review produced to contribute to the report the status of health in the european union: towards a healthier europe," *NDT Plus*, vol. 3, no. 2, pp. 213–224, 2010.

[2] A. R. OMRAN, "The epidemiologic transition: A theory of the epidemiology of population change," *Milbank Quarterly*, vol. 83, no. 4, pp. 731–757, 2005.

[3] W. H. Organization, "Preventing chronic diseases: a vital investment," 2005.

[4] K. Strong, C. Mathers, S. Leeder, and R. Beaglehole, "Preventing chronic diseases: how many lives can we save?" *Lancet*, vol. 366, no. 9496, pp. 1578–1582, 2005.

[5] TeSAN. (2008, Jun.) Chronious official site. [Online]. Available: http://www.chronious.eu

[6] M. Vitacca, L. Bianchi, A. Guerra, C. Fracchia, A. Spanevello, B. Balbi, and S. Scalvini, "Tele-assistance in chronic respiratory failure patients: a randomised clinical trial," *European Respiratory Journal*, vol. 33, no. 2, pp. 411–418, 2009.

[7] Jena official web site. [Online]. Available: http://jena.sourceforge.net

[8] W3C. (2009, Nov.) Owl specifications. [Online]. Available: http://www.w3.org/TR/owl-features/

[9] KDOQI, "KDOQI Clinical Practice Guidelines and Clinical Practice Recommendations for Diabetes and Chronic Kidney Disease." *American journal of kidney diseases : the official journal of the National Kidney Foundation*, vol. 49, no. 2 Suppl 2, Feb. 2007.

[10] K. M. S.C. Punitha and M. Punithavalli, "Article: Impact of ontology based approach on document clustering," *International Journal of Computer Applications*, vol. 22, no. 2, pp. 22–26, May 2011.

[11] J. M. Abasolo and M. Gomez, "Melisa. an ontology-based agent for information retrieval in medicine." in *In: Proceedings of the First International Workshop on the Semantic Web (SemWeb2000*, 2000, pp. 73–82.

[12] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." *Nature genetics*, vol. 25, no. 1, pp. 25–29, May 2000.

[13] O. B. R. Mudunuri and T. Neumuth, in *GI Jahrestagung*, S. Fischer, E. Maehle, and R. Reischuk, Eds., vol. 154. GI, 2009, pp. 1044–1054.

[14] Pubmed official site. [Online]. Available: http://www.pubmed.org

[15] M. B. L. Schneider and D. Koepsell, "Article: Impact of ontology based approach on document clustering," *Frontiers in Artificial Intelligence and Applications*, vol. 229, no. 2, pp. 28–38, May 2011.

[16] GATE. (2008, Jun.) Gate official site. [Online]. Available: http://www.chronious.eu

[17] M. G. J. Brank and D. Mladenic, "A survey of ontology evaluation techniques," in *In In Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005*, 2005.

[18] M. C. A. Gangemi, C. Catenacci and J. Lehmann, "Modelling ontology evaluation and validation," in *Proceedings of the 3rd European Semantic Web Conference (ESWC2006), number 4011 in LNCS, Budva*. Springer, 2006.

[19] C. J. van Rijsbergen, *Information retrieval*. Butterworths, 1979.