Quid, N°. 22, pp. 71-76, Ene-Jun, 2013, ISSN: 1692-343X, Medellín-Colombia

MODELO SEMIAUTOMATICO DE EXTRACCIÓN DE INFORMACIÓN PARA EL MARCADO SEMANTICO DE RESEARCH OBJECT

SEMI-AUTOMATIC EXTRACTION OF INFORMATION MODEL FOR SEMANTIC MARKUP OF RESEARCH OBJECT

Mg. Ingrid Durley Torres
Institución Universitaria Salazar y Herrera, *Centro de Investigación, Grupo GEA.*
Carrera 70 # 52-49,
Medellín, Colombia,
i.torres@iush.edu.co

PhD. Jaime Guzmán-Luna
Universidad Nacional de Colombia, sede Medellín,
Departamento de Ciencias de la Computación y la Decisión,
Grupo de Investigación SINTELWEB.
Carrera 80 No. 65-223,
Medellín, Colombia,
jaguzman@unal.edu.co

Mg. Miguel Alberto Becerra
Institución Universitaria Salazar y Herrera, *Centro de Investigación, Grupo GEA.*
Carrera 70 # 52-49,
Medellín, Colombia,
m.becerra@iush.edu.co

(Recibido e109-02-2013. Aprobado e102-05-2013)

Resumen: este articulo presenta un diagnóstico del estado actual de la investigación en el campo de aplicación de los *Research Object,* donde los actuales avances se reportan en dominios como la bioinformática, la astronomía, la biología y la genética, dejando de manifiesto la necesidad de extender este modelo a otras áreas. Conscientes de este reto, este articulo define el problema y los principales conceptos del dominio de investigación, con el ánimo de proponer un Modelo Semiautomática de Extracción de Información para el Marcado Semántico de *Research Object,* con el fin de obtener los elementos constitutivos de los procesos de investigación expresados por un autor en un documento científico digital.

Palabras claves: aplicaciones educativas, objetos de aprendizaje, web semántica y diversidad social.

Abstract: this paper presents a diagnosis of the current state of Research in the field of application of the Research object, where current developments are reported in domains such as bioinformatics, astronomy, biology and genetics, leaving out the need of extend this model to other areas. Aware of this challenge, this paper defines the problem and the main concepts of the domain of Research, with the aim of proposing a Model Semiautomatic Information Extraction for the Semantic Marked object of Research, in order to obtain the elements of the processes Research expressed by an author in a scientific document digital.

Keywords: educational applications, learning objects, semantic Web and social diversity.

1. INTRODUCCION

Actualmente los autores tienden a escribir muchos artículos o, a lo menos, tantos como sean posibles para conseguir mayor divulgación y referencialidad en el ámbito científico. Siempre que se realiza algún progreso en un determinado tema, un nuevo documento científico es escrito, revisado y publicado de manera sistemática. De esta forma, existe tanta información disponible de los resultados de investigación de un autor, que realizar una trazabilidad de sus procesos de investigación, puede resultar una tarea altamente costosa. Con la sobresaturación de producción científica, mucha información importante es subutilizada o hasta ignorada, ya que la totalidad de artículos, no resulta ni siquiera perceptible a la capacidad humana.

Además de considerar que las estructuras de los datos textuales no es única, razón por la cual la interconexión de palabras no es inmediata o clara para ser comprendida por los humanos o los propios agentes de software, en este campo, se pueden hallar diversos estudios, métodos, metodologías o modelos que reflejan con diversos términos de un experimento un mismo significado, haciendo que la búsqueda de información se vea obstaculizada por él, dramáticamente creciente, volumen de datos y por las disparidades en los pianos técnico y conceptual entre las fuentes de datos individuales.

Desafortunadamente, tanta información ha tejido un abismo entre la producci6n de datos y el consumo y manejo de los mismos. Si un investigador novato o recientemente inmerso en una temática desea replicar, interpretar o adaptar los procesos de investigación de otros, debe primero descifrar el lenguaje usado por cada distinto autor en sus documentos científicos (aunque pertenezcan a la misma temática) y asociarlo a su interés investigativo. Además de identificar, seleccionar e interpretar la información allí descrita, para poder especificar lo que resulta útil o no, a sus intereses. Si a esto le sumamos la cantidad de documentos producidos por todos los autores. La tarea resulta tediosa, altamente costosa en recurso humano y muy demorado.

Esto obliga a pensar, en la posibilidad de habilitar a las máquinas para que realicen la tarea de forma automática, sin la intervención humana. Para ello resulta de especial interés, tratar de modelar el intercambio global de información científica (Hernán, 2010); es decir, unificar las conceptualizaciones mentales y terminológicas, que puedan darse entre un usuario consumidor y uno o varios usuarios productores. Esta es la razón por la cual la propuesta no solo considerar representar, mediante un vocabulario univoco, la representación de un dominio a través de ontologías sino que, a su vez, se pretende potencializar el proceso de intercambio mediante un conjunto de reglas semánticas que ofrezcan una forma de expresar restricciones e inferencias, no explicitadas directamente en las ontologías formuladas. Este proceso semánticamente enriquecido (W3C, 2007), permitir tratar la integración entre fuentes de información heterogéneas procedentes de cada Research Object, sino a su vez, mejorar la eficiencia de las consultas, con tareas de inferencia.

2. MARCO TEORICO

2.1. Research Object

Como se citó previamente, los Research Object, sobrepasan al PDF normal porque pueden además archivar, clasificar e indizar en repositorios semánticos escalables; permitir un acceso avanzado y otorgarles capacidades de recomendación basadas en cl seguimiento y las métricas para evaluar similitudes, la calidad, la estabilidad y la integridad; construir comunidades científicas para compartir colaborativamente, reutilizar y desarrollar los Research Object.

Basándose en la definición de (Verdes-Montenegro 2012), es posible citar textualmente que las propiedades de los Research Object se definen por:

- Reusabilidad: El principio clave de los Research Object está en apoyar el intercambio y la reutilización de los datos, métodos y procesos.
- Configurabilidad: La reutilización, puede involucrar partes que constituyen el Research Object en sí mismo.
- Repetitividad: Debe existir suficiente informaci6n en el Research Object para permitir repetir el estudio, incluso años más tarde.
- Reproducibilidad: Una tercera parte puede comenzar con las mismas entradas y métodos para constatar que los resultados anteriores pueden ser confirmados.
- Temporalidad: los estudios pueden involucrar investigaciones sencillas que tardan milisegundos o prolongados procesos que tarden años.
- Referencialidad: Si los Research Object son para aumentar o reemplazar los métodos de la publicación tradicional, entonces ellos deben ser referenciales o citables.
- Revelable: A los terceros, se les debe permitir auditar los pasos desarrollados en la investigación, con el fin de que sean convencidos de la validez de los resultados.
- Respetuosidad: Representaciones explicitas del origen, procedencia y flujo de la propiedad intelectual.

2.2. Web Semántica

Haciendo use de metadatos y de las reglas de normalización para generar los metadatos de un recurso, se obtienen datos que pueden procesarse de manera más eficiente. Sin embargo, una restricción importante para la recuperaci6n de recursos a través de sistemas automatizados es la incapacidad de

búsquedas semánticas, problema que dentro de la Web sigue causando gran número de respuestas fallidas en los buscadores más potentes, ya que los motores carecen de inteligencia y aún no procesan el significado de las palabras.

Para cubrir esta deficiencia sería necesario utilizar una misma semántica entre los repositorios, quienes buscan en ellos, los autores de contenidos, los que catalogan y los que publican, tarea más que difícil de lograr para una misma comunidad y mucho más difícil entre sectores. Para esto, se está experimentado sobre la Web Semántica, una extensión de la Web actual, en la cual, la información tiene un significado bien definido, permitiendo a ordenadores y a personas trabajar de forma cooperativa (Berners-Lee et al, 2001).

(Garcia, 2004) señala: "La idea de la Web Semántica es tener datos en la Web bien definidos y enlazados de manera que puedan ser usados de forma más efectiva para un descubrimiento, una automatización, una integración y una reutilización entre diferentes aplicaciones. Para ello la Web debe evolucionar, ofreciendo una plataforma accesible que permita que los datos se compartan y se procesen por herramientas automatizadas o personas".

La Web Semántica estructura los recursos disponibles en el Web de forma semántica, para que a través de agentes de software se analicen y se ejecuten procesos principalmente de búsqueda y recuperación.

El desarrollo de la Web Semántica se apoya principalmente en dos tecnologías: XML para el etiquetado de la estructura de un recurso que pueda ser interpretado por una máquina y RDF (Resource Description Framework) [RDF, 2004], para la especificación de metadatos e información sobre el recurso.

Recientemente, para el desarrollo de la Web Semántica, se ha adoptado el concepto de ontologías para encontrar equivalencias en términos comunes que se identifican de forma diferente. Formalmente, una ontología se define como una especificación explicita de una conceptualización (Gruber, 1993), en la práctica se definen vocabularios comunes para compartir información dentro de un determinado dominio, se puede decir que es proceso sofisticado y estructurado para la normalización exhaustiva de la información. Algunos trabajos que citan la Web Semántica, la televisión digital y el ámbito educativo son:

El trabajo de la UPM (Tovar y Bonastre, 2012), se ha dirigido a analizar las tendencias de la televisión digital pero proyectadas a la convergencia de las tecnologías web. Entre las que se destacan: Contenido bajo demanda (Future Networks Media), Explotar nuevas plataformas (Personalización), Nuevos modos de participación (factores sociales), Explotar nuevos factores de datos (Web semántica) y, finalmente, Cerrar el círculo (retroalimentación). Además de ello, el trabajo se concentra en usar la semántica para especificar de manera inequívoca una caracterización de los servicios que oferta la televisión digital. El trabajo, se concentra a un enfoque teórico fuertemente soportado pero, a la fecha, carece de implementaciones debido a que su orientación está determinada a la fase de análisis y su aplicabilidad a futuro.

2.3. Proceso de investigación

De manera común un proceso de investigación es aquel conjunto de actividades y tareas que siguen el método científico (Sampieri 2000), el cual está definido por la especificación de un problema, la formulación de una hipótesis, la definición de unos objetivos (general y especifico), el marco de desarrollo (teórico, jurídico, institucional), la formulación de una metodología, la obtención de resultados y la formulación de unas conclusiones. Para realizar este proceso, comúnmente un investigador debe indagar o escudriñar en la literatura vigente, esta tarea implica a su vez otras sub actividades específicas, también orientadas a la construcción de un proceso de investigación (T. Varela, et al. 2007). En este caso se definen:

Los enfoques del problema. Los análisis realizados del fenómeno estudiado. Las variables dependientes e independientes consideradas como relevantes para el problema, por ejemplo, en un estudio sobre el use de un trombolítico nuevo y otro convencional para el tratamiento del infarto cardiaco. Las variables independientes son los diversos tratamientos; las dependientes: supervivencia, tiempo de hospitalización, etcétera. Los diseños estadísticos utilizados, conocer el tipo de diseño empleado, por ejemplo: series cronológicas con estímulos repetidos, en bloque al azar, cuadro latino, Split- y otras. Las teorías empleadas: Einstein concedió un lugar prominente a la formulación de la estructura matemática (andamiaje) para la creación de su teoría de la relatividad, la novedosa interpretación realizada de la geometría analítica euclidiana y no euclidiana y su inserción en el marco tridimensional con sentido físico... "fue decisiva.

Las técnicas de medición, los resultados obtenidos y las interpretaciones correspondientes.

La aparición de problemas resueltos en momentos anteriores que surgen bajo otras condiciones, por ejemplo, el parasito de la malaria, el Plasmodium falciparum, comenzó a mostrar invulnerabilidad ante los tratamientos y esto llevo a un resurgimiento de la enfermedad. La artemisinina, hasta ahora el medicamento más eficaz, podría dejar de serlo. Cada año mueren por esta causa 2,5 millones de personas.

Las ideas no verificadas de los autores o apreciaciones hipotéticas.

Las coincidencias y contradicciones.

De esta especificación anterior, es que se deberá definir un mecanismo de especificación formal que modele este tipo de procesos.

3. RECONOCIMIENTOS

Los autores agradecen a la Institución Universitaria Salazar y Herrera y a la Universidad Nacional de Colombia por cofinanciar el proyecto de investigación "Modelo Semiautomática de Extracción de Información para el Marcado Semántico de Research Object", convocatoria No. 7.

4. REFERENCIAS

Alper, P, K Belhajjame, and C Goble 2013. Small is Beautiful: Summarizing Scientific Workflows Using Semantic Annotations. Big Data (BigData 2013).

Bechhofer, S, et al. 2010 Research objects: Towards exchange and reuse of digital knowledge: precedings.nature.com.

Bechhofer, S, S Soiland-Reyes, and K Belhajjame 2011. Workflow Lifecycle Management Initial Requirements.

Belhajjame, K, et al. 2012 Workflow-centric Research objects: First class citizens in scholarly discourse: users.ox.ac.uk.

Belhajjame, K, SM Embury, y NW Paton. 2006. On characterising and identifying mismatches in scientific workflows. Data Integration in the Life Sciences. Third International Workshop, DILS 2006, Hinxton, UK, July 20-22, 2006.

Belhajjame, K, SM Embury, and NW Paton, R. Stevens, C. Globe. 2008. Automatic annotation of web services based on workflow definitions. ACM Transactions on the Web, Volume 2 Issue 2, April 2008. Article No. 11. ACM New York, NY, USA

Belhajjame, K, G Vargas-Solar, y G. Collect. 2001. A flexible workflow model for process-oriented applications. Web Information Systems Engineering, 2001. Proceedings of the Second International Conference. 3-6 Dec. 2001. Vol 1. pp 72-80.

Berardi, D. 2005 Automatic Service Composition. Models, Techniques and Tools Doctoral, Ingenier Informatica, University degli Studi di Roma "La Sapienza".

Bohle, S. 2013 "What is E-science and how should it be managed?" Nature.com, Spektrum der Wissenschaft (Scientific American), http://www.scilogs.com/scientific_and_medicallibraries/what-is-e-science-and-how-should-it-be-managed/

Borrajo, D. 2013. Plan Sharing for Multi-Agent Planning. Proceedings of the 1st Workshop on Distributed and Multi-Agent Planning (DMAP 2013), Roma, 2013, pp. 57-66. ICAPS.

Fisher, P, Wolstencroft K, Haines R, Fellows, 2013. The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. http://nar.oxfordjournals.org/content/carly/2013/05/02/nar.gkt328.abstract

David S. The open world. PhD Thesis, Science, Technology, and Society, Cornell University, 2007.

De Roure D., Goble C., y Stevens R. The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows. Future Generation Computer Systems, 25:561-567, 2009

Garijo, D, et al. 2012. Common motifs in scientific workflows: An empirical analysis. E-Science (e-Science), 2012 IEEE 8th International Conference on E-Science

Hettne, KM, et al. 2012 Best Practices for Workflow Design: How to Prevent Workflow Decay: ceur-ws.org.

Lagoze C., Van de Sompel H., Nelson M., Warner S., Sanderson R and Johnston P. (2012), AWeb-based resource model for scholarship 2.0: object reuse & exchange, Concurrency And Computation: Practice And Experience Concurrency Computat.: Pract. Exper. 2012; 24:2221- Published online 8 June 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/cpc.1594

Lee K,, Slattery **0.** Lu R., Tang X., y McCrary V. (2002). The State of the Art and Practice in Digital Preservation. Journal of Research of the National Institute of Standards and Technology. Volume 107, Number 1, January-February.

Missier, P, P Missier, S Dey, K Belhajjame, V Cuevas-Vicenttin. 2013 D-PROV: extending the PROV provenance model with workflow structure: usenix.org. Computing Science, Newcastle University.

Page, K, R Palma, and P Houbowicz 2012 from workflows to Research Objects: an architecture for preserving the semantics of science: linkedscience.org.

PARSE. Insight Project: FP7-2007-223758. 2007 Survey report. http://www.parse-insight.eu/downloads/PARSE-InsightD3-4SurveyReportfinalhq.pdf

Roure, D De, et al. 2009 the myexperiment open repository for scientific workflows: eprints.soton.ac.uk.