Building a Semantic Virtual Museum: from Wiki to Semantic Wiki using Named Entity Recognition

Alain Plantec, Vincent Ribaud, Vasudeva Varma

Building a Semantic Virtual Museum: from Wiki to Semantic Wiki using Named Entity Recognition

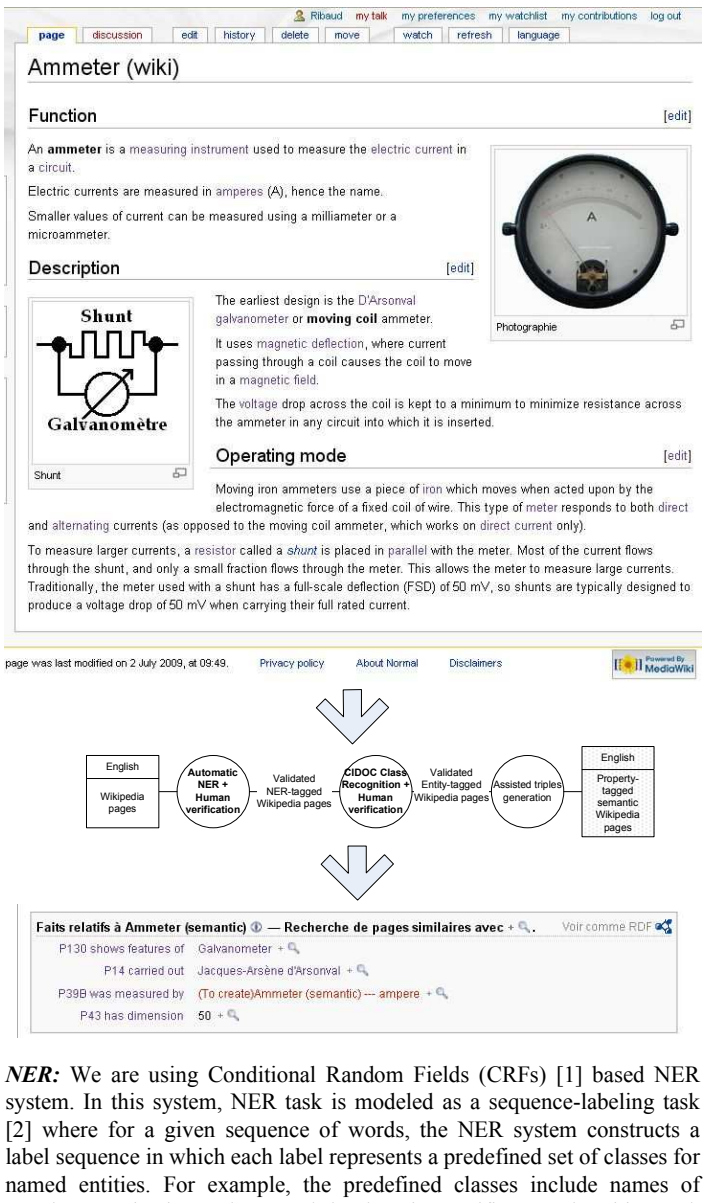
Alain Plantec, Vincent Ribaud, Vasudeva Varma²
Brest University, ²International Institute of Information Technology Hyderabad

Abstract : In this paper, we describe an approach for creating semantic wiki pages from regular wiki pages, in the domain of scientific museums, using information extraction methods in general and named entity recognition in particular. We make use of a domain specific ontology called CIDOC-CRM as a base structure for representing and processing knowledge. We have described major components of the proposed approach and a three-step process involving name entity recognition, identifying domain classes using the ontology and establishing properties for entities in order to generate semantic wiki pages. Our initial evaluation of the prototype shows promising results in terms of enhanced efficiency and time and cost benefits.

A semantic wiki aims to add meaning to the values and the links embedded in wiki pages. Sharing semantics in a community relies on the use of a common ontology of universals: the definition of the domain concepts (classes) and of the properties used to represent and associate concepts. When we are building a semantic wiki to share knowledge about a domain, it requires to identify the particulars (instances) about we have knowledge, then to instanciate classes and properties in order to represent these particulars. Traditional wiki pages contain a lot of unstructured knowledge and our research work aims at providing the end-users with methods and tools that may help in extracting the semantic knowledge from regular wiki pages.

A typical wiki page about an ammeter is represented below. Knowledge about this ammeter may be represented with a list of couples: (property, object) such as (has dimension, 50) or (shows features of, Galvanometer). Semantic MediaWiki represents these couples externally in the Factbox (see below) and internally with RDF triples constructed from the couples (the URI of the wiki page is considered as the subject of all triples embedded in the page). We use the CIDOC CRM, a standardized ontology intended for interpreting cultural heritage data. When we are working on a wiki page, the first step is a human intervention to decide which are the classes where this page is an instance. Then we have to recognize and select the particulars such as 50 or Galvanometer that have to be linked to the particular corresponding to the wiki page, e.g. an Ammeter. The recognition step can be automated through Named Entity Recognition (NER). NER helps also to produce information regarding the type of the particulars (attribute or class), e.g. Number or Instrument, but a human validation is required to select the proper class of the domain ontology. Once the particulars (and their classes) has been properly recognized, the last step is to build a path between the original wiki page and each wiki page associated with one of the recognized particulars. Properties provide a mechanism for expressing relationships between classes. Link between two particulars can be direct if a property between the classes of these particulars exists the CIDOC CRM ontology, or indirect through one or several particulars – existing or to be created. For example, there is no direct association between a Man-Made Thing such as an ammeter and a Dimension such as an ampere; it Is required to create an intermediate instance of the Measurement activity that is linked with the ammeter instance and the ampere instance.

Overview of the translation process



NER: We are using Conditional Random Fields (CRFs) [1] based NER system. In this system, NER task is modeled as a sequence-labeling task [2] where for a given sequence of words, the NER system constructs a label sequence in which each label represents a predefined set of classes for named entities. For example, the predefined classes include names of people, organizations, places and the domain specific named entities such as instruments, part of instrument, instrument material, and measurement. The final label sequence is the one that has highest probability among all the possible label sequences occurring for a given word sequence.

CRM Class Recognition: We have used ICOM/CIDOC Conceptual Reference Model (CIDOC CRM), an ontology intended to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information and an ISO 21117 standard since 2006 as the base structure for organizing information [3]. CIDOC classes are mapped to the tags used by the NER. One-to-one mapping is easy to process, but complex mapping (e.g. Measurement) requires a language to define transformation rules, ideally performed as automated processes.

Assisted triples generation: RDF is a property-centric (rather than record-centric) approach to representation. The domain of a property is used to indicate that a particular property applies to a designated class [4]. The range of a property is used to indicate that the values of a particular property are instances of a designated class [4]. Thanks to NER and class recognition, domain and range of each potential triple is known. A suitable representation of the CIDOC CRM ontology is required to provide the user with the possible choices.

References

[1] Wallach, H.M. 2004. Conditional random fields: An introduction. MS-CIS-04-21, Pennsylvania University.

[2] Lafferty, J., McCallum, A. and Pereira, F. 2001. Conditional random: Probabilistic models for segmenting and labeling sequence data. In 18th Proc. of the Int. Conf. on Machine Learning, 282-289. Morgan Kaufmann, San Francisco.

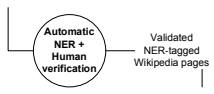
[3] ISO/IEC 21127:2006, "A reference ontology for the interchange of cultural heritage information", ISO, Geneva.

[4] W3C. 2004. RDF/XML Syntax Specification (Revised). http://www.w3.org/TR/rdf-syntax-grammar/

Excerpts of the translation process

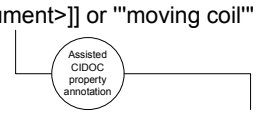
This translation process is performed in three steps: named entity recognition (NER), CIDOC-CRM class recognition, and CIDOC-CRM property disambiguation. The two former steps are automatic but require human validation to ensure that named entities or CRM classes were correctly recognized. The last step is a computer-aided process.

An '''ammeter''' is a [[measuring instrument]] used to measure the [[electric current]] in [[ampere]]s (A), hence the name.
…
The earliest design is the [[Jacques-Arsène d'Arsonval|D'Arsonval]] [[galvanometer]] or '''moving coil''' ammeter.



An '''<Instrument>ammeter</Instrument> is a [[measuring instrument]] used to measure the [[electric current]] in a [[Electrical circuit|circuit]]. Electric currents are measured in [[<Measure>ampere</Measure>]]s (A), hence the name.
...
The earliest design is the [[<Person>Jacques-Arsène d'Arsonval</Person>|<Person>D'Arsonval</person>|]] [[<Instrument>galvanometer</Instrument>]] or '''moving coil''' ammeter.

An '''<E101 Instrument>ammeter</E101 Instrument> is a [[measuring instrument]] used to measure the [[electric current]] in a [[Electrical circuit|circuit]]. Electric currents are measured in [[<E16 Measurement><E54 Dimension>ampere</E54 Dimension></E16 Measurement>]]s (A), hence the name.
…
The earliest design is the [[<E21 Person>Jacques-Arsène d'Arsonval</E21 Person>|D'Arsonval]] [[<E101 Instrument>galvanometer</E101 Instrument>]] or '''moving coil''' ammeter.



An '''ammeter''' is a [[measuring instrument]] used to measure the [[electric current]] in a [[Electrical circuit|circuit]]. Electric currents are measured in [[ampere]]s (A), hence the name.
...
The earliest design is the [[P14 carried out::Jacques-Arsène d'Arsonval|D'Arsonval]] [[P130 shows features of::galvanometer]] or '''moving coil''' ammeter.

[[P39B was measured by::(To create)Ammeter (semantic) --- ampere|]]
…
[[Category:E101 Instrument]]

Wiki pages are composites

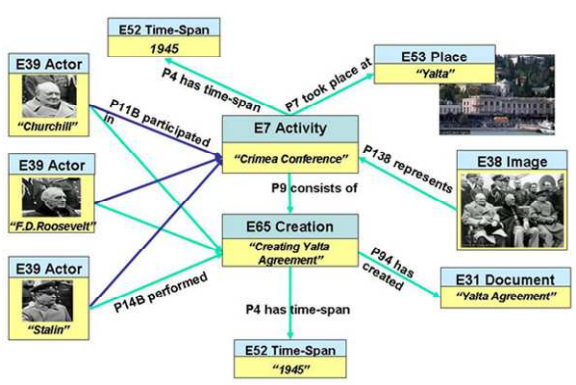
Yalta Conference

Photograph from the Army Signal Corps Collection in the U.S. National Archives.

The Crimea Conference, known as the Yalta Conference took place in 1945 at Yalta. The picture left represents three famous heads of government of the United States, the United Kingdom, and the Soviet Union - President Franklin D. Roosevelt, Prime Minister Winston Churchill, and Josef Stalin, from left to right -.
This conference led to the creation of the Yalta agreements, materialized by the Protocol of Proceedings at the Yalta Conference (11 February 1945) and signed by these three heads of government.

This graph represents the decomposition of the composite « Yalta Conference ». If we assume that Actors – such as Roosevelt or Churchill, Things – such as the document Yalta agreements, Images, Time-Span and Places are particulars that exists on their own in the wiki, the whole Event called « Yalta Conference » is composed of 2 Activities (the conference itself and the creation of the Yalta agreements). Hence, any of the components of the composite may be the subject or the object of the RDF triples that can be extracted from the whole page.

RDF provides constructs to deal with groups of things: containers and collections, but Semantic MediaWiki (SMW) features do not include composites processing. Current work is about some additional paradigm for SMW that are required to support (semantic) composition. The problem of creating a composite differs from the problem of referring to a composite. When an user decides to transform a page as a composite, s/he is probably able to decide which triples are related to which parts of the composite. But existing links to this page may be re-processed to reattribute the incoming triples to the right part of the composite.



Our approach assumes that all the knowledge embedded in a wiki page is related to the same particular – acting as the subject of all generated RDF triples. But, most pages are composites that combines several entities into a collective entity that can be referenced as if it were atomic.

A typical use of composites in a wiki is to see a page as a collection of subsections and images that can be edited (and referenced) independently. The page can also be manipulated, e.g. read, referenced or printed, as a single entity.