Building a Semantic Virtual Museum: from Wiki to Semantic Wiki using Named Entity Recognition
Alain Plantec, Vincent Ribaud, Vasudeva Varma

To cite this version:

HAL Id: hal-02912811
https://hal.univ-brest.fr/hal-02912811
Submitted on 24 Aug 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
A semantic wiki aims to add meaning to the values and the links embedded in wiki pages. Shared 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 which we have knowledge, then to instantiate 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 (dimension, 50) or (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 cultural heritage data and if an instance is an instance. Then we have to recognize 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 of Instruments (ameter) or a list of names required to specify the class of the particular in the regular 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

**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]{A} (A), hence the name.

The earliest design is the [[Jacques-Arsène d'Arsonval|D'Arsonval]] [[galvanometer]] or "moving coil" ammeter.

The earliest design is the [[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 [Measure>ampere</Measure>](A), hence the name.

The earliest design is the [[<Person>Jacques-Arsène d'Arsonval</Person>]] [[<Instrument>galvanometer</Instrument>]] or "moving coil" ammeter.

The earliest design is the [[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 [Measure>ampere</Measure>](A), hence the name. The earliest design is the [[<Person>Jacques-Arsène d'Arsonval</Person>]] [[<Instrument>galvanometer</Instrument>]] or "moving coil" ammeter.

[[P39B was measured by::(To create)Ammeter (semantic) --- ampere| ]] (A), hence the name.

The earliest design is the [[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 [Measure>ampere</Measure>](A), hence the name.

The earliest design is the [[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 [Measure>ampere</Measure>](A), hence the name.

Wiki pages are composites

**Yalta Conference**

Photographed from the Army Film and Photo Co. 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 Joseph 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 (11th 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 and Places are particulars that exist on their own in the wiki, the whole Event « Yalta Conference » is composed of 2 Activities (the creation of the Yalta conf. 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 document.

RDF provides constructs to deal with groups of things: containers and collections, 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.

We may not be able to decide which triples are related to which parts of the composite. But existing links to this page may be re-referenced to re-represent 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 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.