# Building and Querying Semantic Layers for Web Archives

Vaibhav Kasturia, Pavlos Fafalios, Helge Holzmann and Wolfgang Nejdl

## Problem

Despite the increasing number of web archives worldwide, the absence of efficient and meaningful exploration methods still remains a major hurdle in the way of turning web archives into a usable and useful source of information.

For a bit more complex information needs, like *exploratory search needs*, keyword-based search in web archives leads to ineffective interactions and poor results. Thus, there is the need to go beyond keyword-based search and support more advanced information seeking strategies.

## Contributions

We make the following contributions:

1. We introduce a simple but flexible RDF/S data model, called *Open Web Archive*,
2. We detail the process of constructing semantic layers and we present an open source and distributed framework, called *ArchiveSpark2Triples*, that facilitates their efficient construction.
3. We present (and make publicly available) three semantic layers for three different types of web archives: one for a *versioned web archive*, one for a *non-versioned news archive*, and one for a *social-media archive*.
4. We present the results of a comparative evaluation using a set of 20 information needs of exploratory nature.

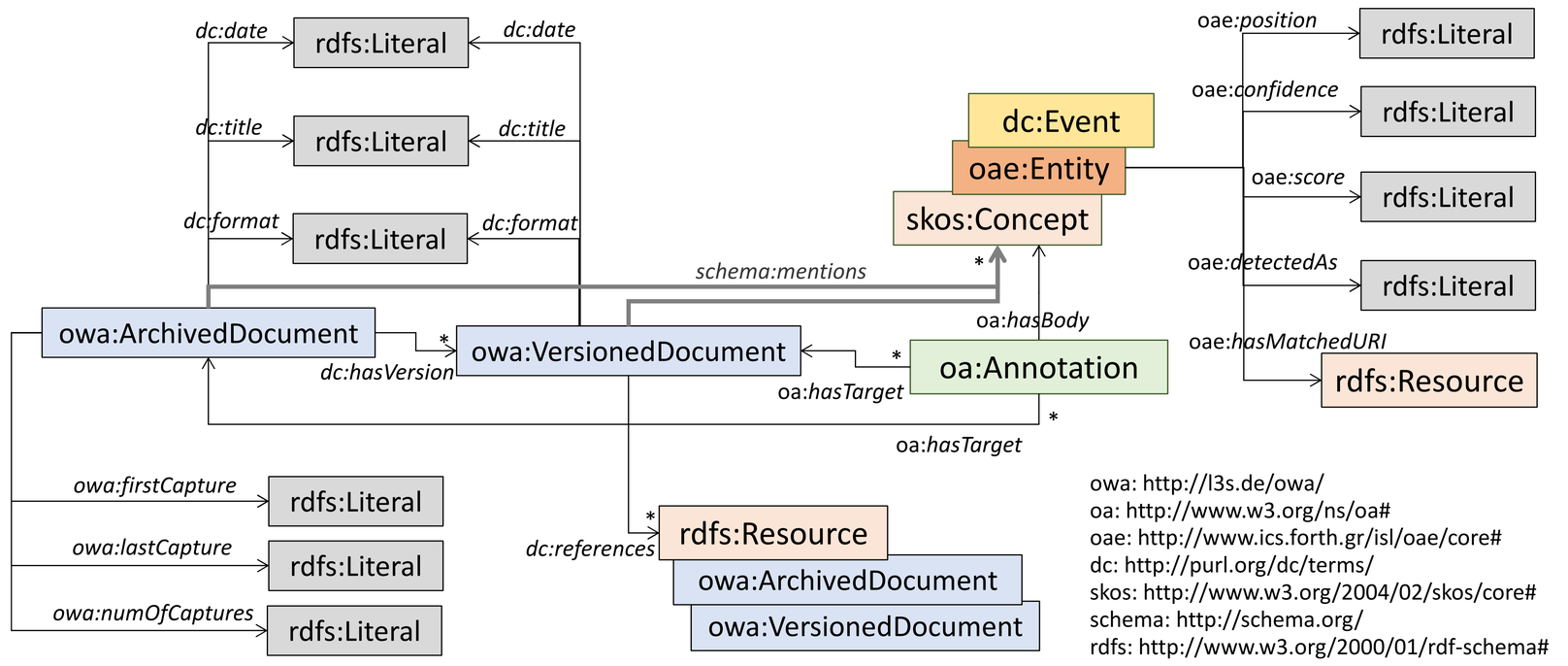
## Results

We have introduced a model and a framework for describing and publishing metadata and semantic information about web archives. The constructed *semantic layers* allow:

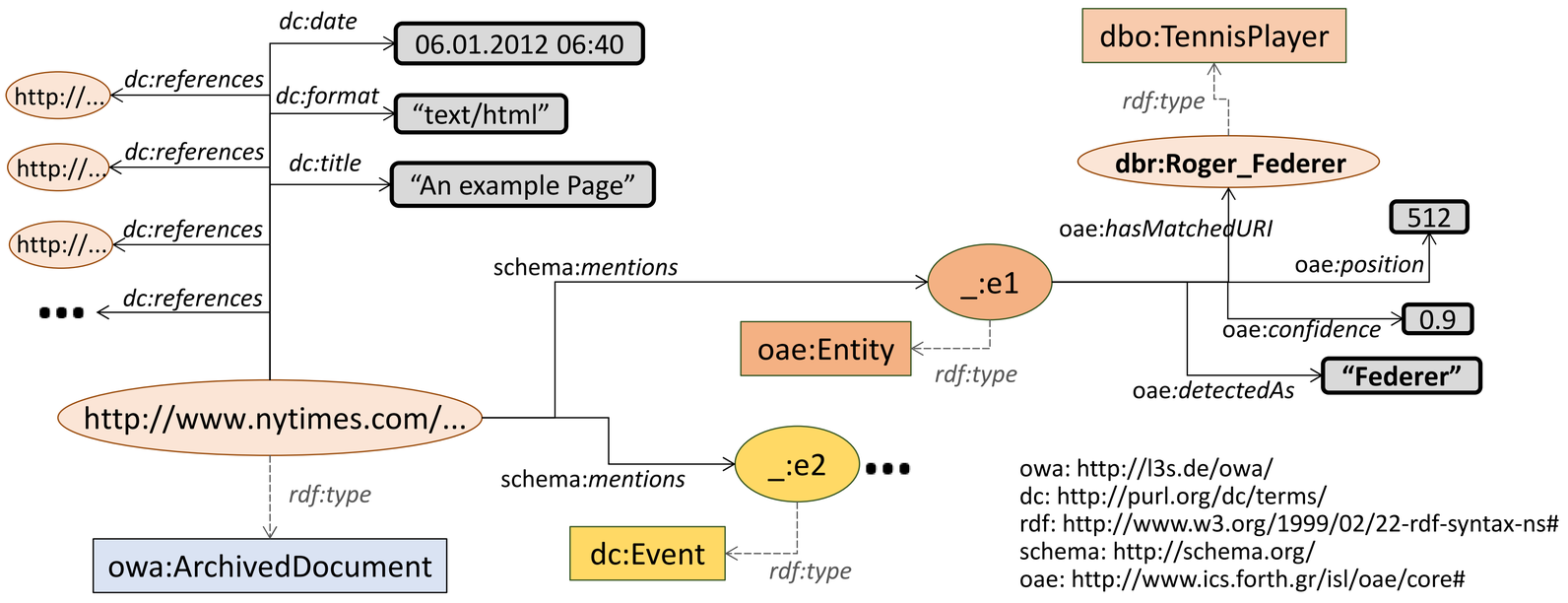
1. exploring web archives in a more advanced way based on entities, events and concepts extracted from the archived documents and linked with web resources.
2. integrating information coming from multiple knowledge bases and semantic layers.
3. inferring new knowledge that is very laborious to derive otherwise.
4. coping with common problems when exploring web archives like temporal reference variants and multilinguality.
5. making the contents of web archives machine understandable, thereby enabling their direct exploitation by other systems and tools.

The results showed that semantic layers can answer complex information needs that keyword-based search systems fail to satisfy.

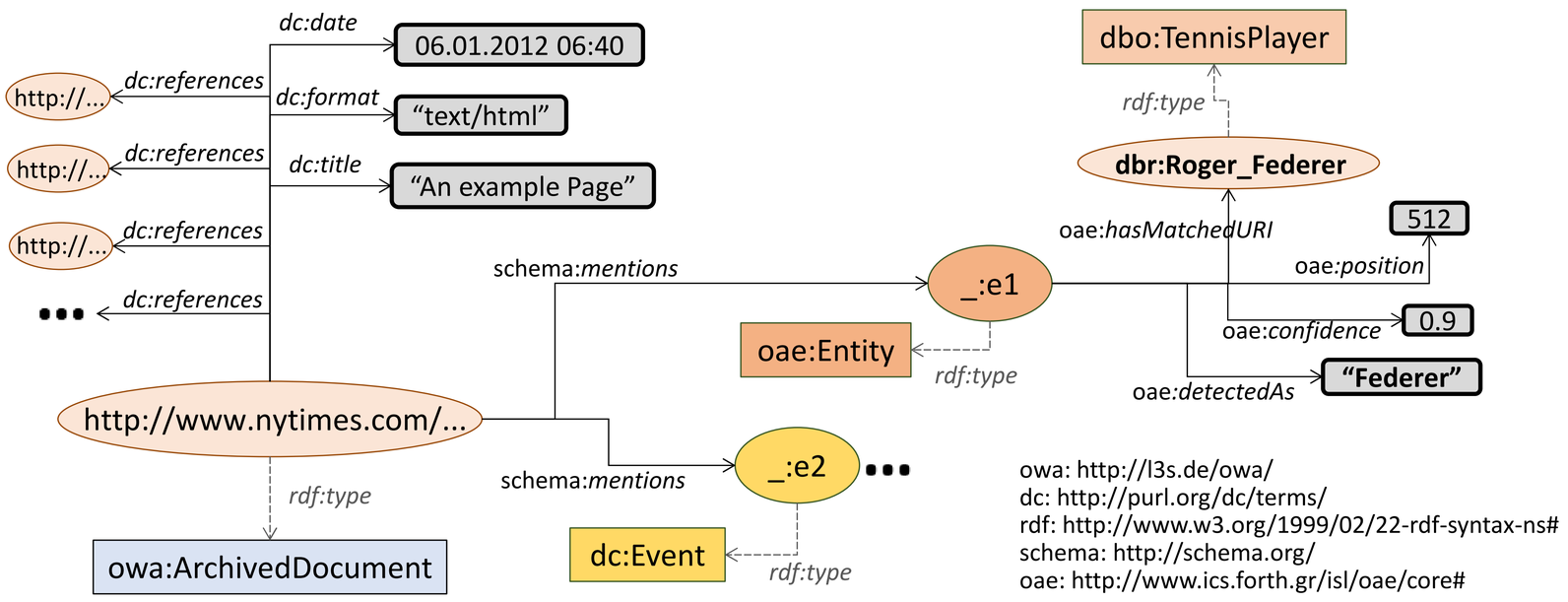
## Data Model




The *Open Web Archive* data model

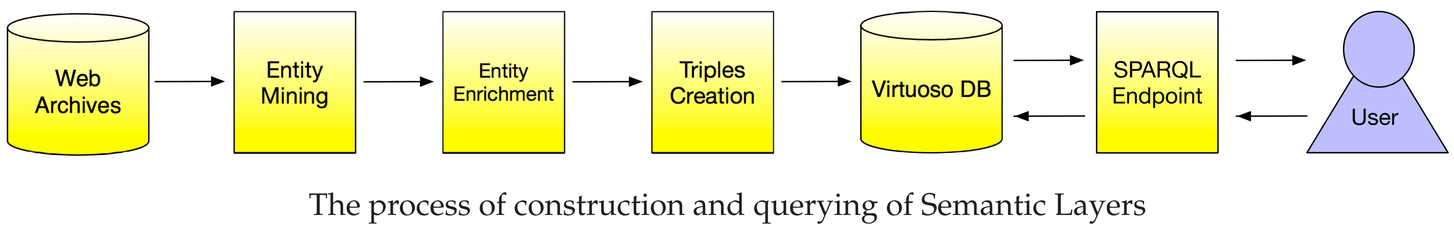



Describing an archived article (non-versioned) using the *Open Web Archive* data model




Describing an archived web page (versioned) using the *Open Web Archive* data model

## Framework




The process of construction and querying of Semantic Layers

## Future Work

1. Regarding future work and research, user-friendly interfaces should be developed on top of semantic layers for allowing end-users to easily and efficiently explore web archives.
2. Another interesting direction is to study approaches for ranking the results returned by SPARQL queries.

## Source Code

The Semantic Layers are publicly available at: `http://l3s.de/owa/semanticlayers/`

The dataset used for evaluation is available at: `http://l3s.de/owa/semanticlayers/SemLayerEval.zip`