Identity of Resources and Entities on the Web

Valentina Presutti, ISTC National Research Council (CNR), Italy
Aldo Gangemi, ISTC National Research Council (CNR), Italy

ABSTRACT

One of the main strengths of the Web is that it allows any party of its global community to share information with any other party. This goal has been achieved by making use of a unique and uniform mechanism of identification, the uniform resource identifiers (URI). Although URIs succeed when used for retrieving resources on the Web, their suitability for identifying any kind of thing, for example, resources that are not on the Web, is not guaranteed. In this article we investigate the meaning of the identity of a Web resource, and how the current situation, as well as existing and possible future improvements, can be modeled and implemented on the Web. In particular, we propose an ontology, IRE, that provides a formal way to model both the problem and the solution spaces. IRE describes the concept of resource from the viewpoint of the Web, by reusing an ontology of information objects, built on top of DOLCE+ and its extensions. In particular, we formalize the concept of Web resource, as distinguished from the concept of a generic entity, and how those and other concepts are related, for example, by different proxy for relations. Based on the analysis formalized in IRE, we propose a formal pattern for modeling and comparing different solutions to the problems of the identity of resources.

Keywords: DOLCE+; information objects; IRE model; uniform resource identifiers (URI)

INTRODUCTION

The Web is an information space realized by computationally accessible resources, each embedding some information, which is encoded in some language, and expresses some meaning. One of the successful achievements of the Web is allowing different parties of its global communities to share information (Jacobs & Walsh, 2004). Typically, typing an address in a Web browser is enough to visualize or download an object, the meaning of which can be then understood by a human agent. The Web address is a uniform resource identifier, a URI (Berners-Lee, Fielding, & Masinter, 2005). The URI mechanism is key to the Web success. However, another ambitious goal of the Web is that of referencing things in general. For example, consider the World Wide Web Consortium (W3C)’s URI http://www.w3.org: it should be possible to distinguish (on the Web)
the reference to the organization from that to its Web site.

The simple association of a URI to a thing or real world entity is very powerful. On one hand, it has already demonstrated its effectiveness with regard to the identification of objects that are accessible through the Web, for example, Web pages. On the other hand, there is no complete consensus on how to manage identification of things that are not on the Web. Reducing the ambiguity of identifying the entities a Web resource refers to is essential for information sharing, interoperability, and reasoning on the Web (Berners-Lee et al., 2006). In order to propose solutions to this issue, it is crucial to analyze and properly describe the problem space.

The problem space can be expressed in terms of the impact that identification of (generalized) resources has on the Web. In this article we analyze the state of art related to this problem, and from this analysis we show how five distinct issues emerge. We propose that in order to describe these issues and to compare the respective solutions, we need to analyze the reason why a URI can be associated with an entity. We carry out this analysis based on an ontology called identity of resources and entities on the Web (IRE).

IRE focuses on four main classes: URI, Web resource, information object, and entity, which encompass the things in the domain of discourse of the Web referencing problem.

Once the problem domain has been analyzed, the solution domain can be approached. We discuss how the current evolution of Web science from the confluence of the Web, the Web 2.0, and the Semantic Web has affected the solution domain. We also consider some proposed and envisaged solutions, and discuss them in terms of IRE.

The rest of the article is organized into sections as follows: “history” tells a story about the existing literature on the problem of identifying a Web resource. “Issues in the problem space” discusses how the problem of resource identification impacts on the Web. “The IRE Metamodel” informally presents the IRE ontology. We then deal with the “solution space,” and we also present an extension of IRE in order to represent it. “Conclusion and remarks” summarizes the main arguments presented. Finally, the appendix contains a first-order logic formalization of IRE. The OWL version of IRE can be downloaded from http://wiki.loacnr.it/index.php/LoaWiki:IRE.

HISTORY

The identification of resources is an important task to use them on the Web (Berners-Lee et al., 2006). Currently, there is a diffuse feeling that resource identification procedures suffer from a lack of consensus about how to handle them. This lack of consensus partially finds its root from normative documents where the concept “resource” has been defined in the context of the Web. However there are also other motivations underlying the identification problem, which we discuss in this article.

The term “resource” is generally used for all things that might be identified by a URI (Jacobs et al., 2004). In the literature, we find several definitions for the term “resource” used in the context of world wide Web. In particular we quote here three normative documents, IETF RFC 2396 (Berners-Lee et al., 1998), IETF RFC 3986 (Berners-Lee et al., 2005), the W3C’s “Architecture of the World Wide Web” (Jacobs et al., 2004) and discuss the way and consequences of the definition they provide for “resource.” In IETF RFC 2396 the concept of resource is defined as follows (Berners-Lee et al., 1998):

A resource can be anything that has identity. Familiar examples include an electronic document, an image, a service (e.g., “today’s weather report for Los Angeles”), and a collection of other resources. Not all resources are network retrievable: human beings, corporations, and bound books in a library can also be considered resources. The resource is the conceptual mapping to an entity or set of entities, not necessarily the entity that corresponds to that mapping at any particular instance in time. Thus a resource can remain constant even when