



Designing data for the open world of the Web

Tom Baker

Resource Description Framework (RDF) is the grammar for a language of data. Uniform Resource Identifier (URI)s are the words of that language. As in natural language, these words (i.e., the URIs) belong to grammatical categories. RDF properties (such as "isReferencedBy") function a bit like verbs, RDF classes like nouns. As in natural languages, where utterances are meaningful only if they follow a sentence grammar, RDF statements follow a simple and consistent three-part grammar of subject, predicate, and object. Analogously to paragraphs, RDF statements are aggregated into RDF graphs.

Aside from being words in the language of data, URIs double as footnotes. As footnotes they indicate the maintenance responsibility for words by way of ownership of the domain names under which the URIs were coined, as recorded in the globally managed Domain Name Service (DNS). Inasmuch the URIs of words lead to documentation of official definitions, the web itself provides the language of data with its dictionary. The fifteen elements of Dublin Core have been likened to a "pidgin" – a lexicon of generic predicates good enough for the sort of rudimentary but serviceable communication that occurs between speakers of different languages. Just as pidgins are inadequate for more subtle or differentiated expression,



a healthy ecosystem of RDF vocabularies needs to include more specialized vocabularies for use by social or scholarly communities of discourse among themselves. RDF is a language designed by humans for processing by machines. The RDF language – the grammar together with available RDF vocabularies – does not itself solve the difficulties of human communication any more than the prevalence of English guarantees world understanding. However, RDF does support the process of connecting dots – of creating “knowledge” – by providing a linguistic basis for expressing and linking data. Just as English as a second language provides a basis for communication among non-native English speakers, RDF provides a common second language into which local data formats can be translated and exposed. Just as English is useful without being the best of all possible grammars, RDF happens to be what we currently have – the only general-purpose language for data with any traction. But just as English grammar follows deep linguistic structures determined by the human capacity for language, it is likely that RDF, if re-invented, would end up strongly resembling what we currently have. Aside from supporting data interchange in the here and now, RDF provides a response to the ongoing and inevitable obsolescence of computer applications and customized data formats by expressing knowledge using a well-understood grammar and citing publicly documented vocabularies and resource URIs. In this sense, it supports data that does not require additional out-of-band information for its interpretation, i.e., data that “speaks for itself”. This assumes, of course, that our cultural memory institutions will deploy robust methods for preserving the parts of the Web where the underlying RDF vocabularies and resource identifiers are documented. We are in the midst of a rapid shift from a world in which information was predominantly print-based to one in which it is predominantly digital. The scale and speed of transformation virtually

guarantees that any computer applications and user interfaces we use today will at some point, probably soon, be superseded. Data that cannot speak for itself will be more vulnerable to becoming irrelevant.

Not only is data expected to be linkable in the present, but we hope they will be remain intelligible in the future. In 2012, to put information into ad-hoc data formats in the absence of well-defined interpretations as RDF triples is like making statements without grammar. Creating data without URIs is like writing without proper footnotes. This is okay for information with a short shelf life – i.e., most information – but information of lasting cultural significance deserves better. Cultural memory institutions live by the ethos of scholarship, by which things like good grammar and proper footnotes should really matter. The language of RDF represents the application of that ethos to data itself.

TOM BAKER, Dublin Core Metadata Initiative.

thomas.baker@bi.fhg.de

Baker, T. "Designing data for the open world of the Web". *JLIS.it*. Vol. 4, n. 1 (Gennaio/January 2013): Art: #6308. DOI: [10.4403/jlis.it-6308](https://doi.org/10.4403/jlis.it-6308). Web.

ABSTRACT: The domain name system of the world wide web provides a managed space of globally unique identifiers for web pages – Uniform Resource Identifiers, or URIs. URIs can also be used to name things – specifically, to name things in the world ("people," a "books," or "Nelson Mandela"); to name concepts used to describe those things ("Renaissance Sculpture" or "Lyme Disease"); and to name relationships between things (this book "was translated by" that person). Because URIs, used as names, are globally unique, they serve to anchor the strands in "webs of meaning" ("semantic web"). Each strand of the web is a statement following a grammar, the Resource Description Framework (RDF), which uses URIs as its words. Each RDF statement expresses a simple idea – "Dante wrote *L'Inferno*" or "Dante was born in Florence" – which, taken together, can express complex webs of relationships. Expressing data as statements makes it easy to integrate data across many different sources ("linked data"). The opportunity for cultural heritage lies in translating the traditions of resource description into the language of URIs so that its descriptions of Works, Items, Subject Headings, and People can serve as central hubs in growing webs of linked data.

KEYWORDS: Semantic web; Library linked data; URI

Submitted: 2012-06-10

Accepted: 2012-08-31

Published: 2013-01-15

