Data.dcs: From Legacy to Linked Data

Matthew Rowe

OAK Group, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, S1 4DP Sheffield, United Kingdom
m.rowe@dcs.shef.ac.uk

Abstract. The University of Sheffield’s Department of Computer Science provides a web site containing legacy data describing members of the department, their publications and research groups. The Data.dcs project is designed to produce linked data from this legacy data in order to provide an information source for linked data consumers. This paper describes the initial developments of the project and presents the approach through which the first linked dataset was produced.

1 Introduction

The University of Sheffield’s Department of Computer Science (DCS) provides a web site containing legacy data describing members of the department, their publications and research groups. This legacy data is contained with HTML documents - describing research groups and their members - and RSS feeds - describing publications. As a result information is not provided in a coherent and consistent manner which inhibits the ability of end users to find information about the department or query the data. The Data.dcs project addresses this problem by providing linked data, describing the DCS. This paper presents the project’s preliminary work and documents the approach which was implemented to produce the first linked dataset.

2 Triplification

The goal of our approach is to produce linked data describing DCS members, their publications and the research groups which they are members of. For DCS members the DCS web site is crawled for all HTML documents. For publications the RSS feed containing all the publications published by members of the DCS is accessed. To extract legacy data, context windows are derived from the HTML documents and the RSS feed - where one context window contains either information about a person or publication - and Hidden Markov Models (HMMs) are used to extract person or publication information from the windows. From this extracted information triples are built: creating instances of foaf:Person for DCS members, and bibtex:Entry for publications, and assigning the instances the extracted legacy data (i.e. publication title and author). Legacy data within the DCS changes frequently as new people join research groups, old members
leave and new publications are added. To provide a consistent, and up-to-date, linked dataset we create a provenance information graph for every legacy data fragment extracted from the DCS web site. We use the Provenance Vocabulary (PRV) \(^1\) to capture the time and date when the data fragment was accessed, and the service which accessed the fragment. We have extended this vocabulary to include an XPath expression to the data fragment within a given HTML document. Using provenance information enables our triplification process to be deployed as a service which monitors the DCS web site, and produces new triples by reasoning over the provenance information graph attributed to a given data fragment.

3 Building Linked Data

Following the triplification of legacy data into a machine-processable form we build our linked dataset. First we construct instances of foaf:Group for each of research group within the DCS. Using each group’s personnel listing page, we identify all the group members. From this member list we identify coreferring references within the source dataset using SPARQL rules which identify equivalent instances based on inverse functional properties (i.e. same email, homepage) and co-occurrence (i.e. colleagues appearing in the same page). URIs are minted for each group (i.e. \(<http://data.dcs.shef.ac.uk/group/oak>\)) and each member. Publications are then attributed to group members using SPARQL Rules to detect authorship by identifying colleagues as coauthors. The dataset is then linked into the web of linked data by querying DBLP, and using SPARQL rules based on colleague coauthorship to identify publications written by DCS members. The members are also associated with their equivalent instances in the DBLP dataset using an owl:sameAs relation. We have exposed our linked dataset using the "Rules of Linked Data"\(^2\) so that every research group, group member and publication within the DCS has a URI. We have published our dataset using Recipe 2 from "Guidelines to publishing RDF vocabularies"\(^3\) such that each URI is dereferenceable and returns an RDF/XML response when looked up. The first release of our dataset can be found on the Data.dcs web site\(^4\).

4 Conclusions

This poster presents the initial approach deployed by the Data.dcs project which consists of two stages: triplification - extracting legacy data and converting the data into triples - and linked data compilation - building a linked dataset from the triples. Future work will concentrate on the latter of these stages, focussing on statistical methods for URI disambiguation for linking into the web of linked data.

\(^1\) http://purl.org/net/provenance/ns#
\(^2\) http://www.w3.org/DesignIssues/LinkedData.html
\(^3\) http://www.w3.org/TR/swbp-vocab-pub/
\(^4\) http://data.dcs.shef.ac.uk