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