Maintaining an Online Bibliographical Database: The Problem of Data Quality

Michael Ley*, Patrick Reuther*

*Department for Databases and Information Systems, University of Trier, Germany
{ley,reuther}@uni-trier.de
http://dbis.uni-trier.de http://dblp.uni-trier.de

Abstract. CiteSeer and Google-Scholar are huge digital libraries which provide access to (computer-)science publications. Both collections are operated like specialized search engines, they crawl the web with little human intervention and analyse the documents to classify them and to extract some metadata from the full texts. On the other hand there are traditional bibliographic data bases like INSPEC for engineering and PubMed for medicine. For the field of computer science the DBLP service evolved from a small specialized bibliography to a digital library covering most subfields of computer science. The collections of the second group are maintained with massive human effort. On the long term this investment is only justified if data quality of the manually maintained collections remains much higher than that of the search engine style collections. In this paper we discuss management and algorithmic issues of data quality. We focus on the special problem of person names.

1 Introduction

In most scientific fields the amount of publications is growing exponentially. The primary purpose of scientific publications is to document and communicate new insights and new results. On the personal level publishing is a sort of collecting credit points for the CV. On the institutional level there is an increasing demand to evaluate scientists and departments by bibliometric measures, which hopefully consider the quality of the work. All aspects require reliable collection, organization and access to publications. In the age of paper this infrastructure was provided by publishers and libraries. The internet, however, enabled new players to offer services. Consequently many specialized internet portals became important for scientific communities. Search engines like Google(-Scholar) or CiteSeer, centralized archives like arXic.org/CoRR and a huge number of personal and/or department web servers make it very easy to communicate scientific material.

The old players — publishers, learned societies, libraries, database producers etc. — face these new competitors by building large digital libraries like ScienceDirect (Elsevier), SpringerLink, ACM Digital Library or Xplore (IEEE) in the field of computer science.

DBLP (*Digital Bibliography & Library Project*) (Ley, 2002) is an internet "newcomer" that started service in 1993. The DBLP service evolved from a small bibliography specialized to *database systems* and *logic programming* to a digital library covering most subfields