Content Profiling for Preservation: Improving Scale, Depth and Quality

Artur Kulmukhametov[1] and Christoph Becker[1,2]

[1] Information and Software Engineering Group
Vienna University of Technology, Austria
http://www.ifs.tuwien.ac.at/dp
[2] Faculty of Information
University of Toronto, Canada
http://ischool.utoronto.ca/christoph-becker

Abstract. Content profiling in digital preservation is a crucial step that enables controlled management of content over time. However, large-scale profiling is facing a set of challenges. As data grows and gets more diverse, the only option to control it is to combine outputs of multiple characterization tools to cover the varieties of formats and extract features of interest. This cooperation of tools introduces conflicting measures and poses challenges on data quality. Sparsity and labeling conflicts make it difficult or impossible to partition, sample and analyze large metadata sets of a content profile. Without this, however, it is virtually impossible to manage heterogeneous collections reliably over time.

In this paper, we present the content profiling tool C3PO, which includes rule-based techniques and heuristics designed for conflict reduction. We conduct a set of experiments in which we assess the effect of creating such a mechanisms and rule set on the quality and effectiveness of content profiling. The results show the potential of simple conflict reduction rules to strongly improve data quality of content profiling for analysis and decision support.

Keywords: Digital Preservation, Characterization, Content Profiling, Conflict Reduction.

1 Introduction

A crucial starting point for any digital curation process is a full awareness of the set of objects at hand and an assessment of their alignment with the needs of the users, the capabilities of the organization and the evolving context of the digital ecosystem. For digital preservation, such an assessment strongly relies on mechanisms such as characterization and property extraction tools and leverages content profiling to achieve a comprehensive overview on the data held in a repository. A full awareness of data is achievable through running rich in-depth characterization which provides a nuanced view on the diversity of collections, identify risks or help understanding evolution of features. In particular, characterization enables focused preservation planning.

K. Tuamsuk et al. (Eds.): ICADL 2014, LNCS 8839, pp. 1–11, 2014.

Despite a variety of characterization tools available nowadays, there is no single tool that would cover all data types and their properties [13]. In such situations, combining several tools is the only practical approach to cover the heterogeneity of digital artefacts. This raises a new set of challenges:

Depth. Which tools can we use to address this heterogeneity, and how can we combine their output?

Quality. How do we deal with conflicting values? How can we leverage additional tools to improve the quality rather than report conflicts?

Scale. How can we effectively analyze the substantial amount of metadata that is produced when combining multiple tool results?

This paper addresses these challenges and in particular focuses on the improvement of data quality to enable in-depth profiling at scale. We describe the scalable content profiling tool C3PO and introduce a set of improvements, including a mechanism for extensible pre-processing based on a stateless rule engine as part of the gathering process that populates the database of the profiling tool. We describe an experiment on a publicly available large data set, present the resulting rule set, and assess the effect of creating such mechanisms and rule set on the quality and effectiveness of content profiling. The results demonstrate that this is a very cost-effective and robust mechanism for improving the quality of content profiles, which in turn can improve the quality of curation and preservation decisions substantially.

The remainder of this paper is organized as follows: Section 2 gives an overview of related work in characterization and content profiling. Section 3 discusses challenges during content analysis and describes the contribution to address these. Experimentation and results are presented in Section 4. Finally, Section 5 provides conclusions and a short outlook on future work.

2 Characterization and Content Profiling

Characterization is a complex process of taking measures that result in characteristics describing the properties of the content in focus. More specifically, according to [1] we can distinguish 3 aspects of characterization: *identification* of a data structure of a content by file format name and file format version, *format validation* by checking a data structure of a digital object against its format specification and *feature extraction* from characteristics of interest of the content. There is no need to consider all 3 modes of characterization only to obtain general knowledge such as the format name or version. However, deeper characterization will reveal much more detailed insight into the features and risks of a given set of digital objects.

The question arises how many properties should be considered for characterization. There are different view points on this question. From one side, it is possible to select a minimum of properties, a lowest common denominator that can be applied across any type of content. An example of such an approach may be to restrict characterization to producing format profiles [4], which are created by characterization of 2 features - a file format name and a format version.