Large-scale content profiling for preservation analysis

Petar Petrov
Vienna University of Technology
Vienna, Austria
petrov@ifs.tuwien.ac.at

Christoph Becker
Vienna University of Technology
Vienna, Austria
becker@ifs.tuwien.ac.at

ABSTRACT

The starting point of any operational endeavor to preserve digital content is gaining a deep understanding of the characteristics of the objects. Systematic analysis of digital object sets and the identification of sample objects that are representative of a collection are critical steps towards preservation operations and a fundamental enabler for successful preservation planning: Without a full understanding of the properties and peculiarities of the content at hand, informed decisions and effective actions cannot be taken. This article presents a software tool prototype that is able to profile large sets of meta data in a scalable fashion and provide deeper insight into the digital collection at hand.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Information Storage and Retrieval—*Content Analysis and Indexing*; H.3.7 [**Information Systems**]: Digital Libraries—*Collection*

Keywords

Digital Preservation, Preservation Planning, Content Profiling, Characterization, Stratification, Scalability

1. INTRODUCTION

Digital preservation is increasingly becoming relevant for large-scale collections, due to the increase of born-digital material in recent years. Content holding institutions commonly deploy digital content repositories that provide content management facilities and support for large data volumes. However, there is often no comprehensive overview about the detailed types of data contained. Apart from general information such as the number of objects, formats, mime-types and size, there is often a lack of deeper knowledge about the digital objects at hand. Although the meta data of each object can be produced automatically, currently there is no easy way of obtaining a deeper insight into digital collections. The starting point of digital preservation is the deep understanding of the objects at risk. Preservation Planning provides methods for effective decision making that are increasingly supported by automated tools and work-flows. It relies on descriptive information about the objects and carefully chosen samples [5] to conduct controlled experiments and analysis for the purpose of decision support and documentation. A key part of a preservation plan is the description of the collection [1]. This content profile does not describe the information held in the objects, nor their specific domain purpose (blueprints, newspaper articles, government emails, etc.). It serves a far more specific purpose: It aggregates and combines the meta data of the objects in order to give a better overview to the planner and help her understand the implications of the chosen preservation alternatives. Unfortunately, such meta data profiling is often neglected because of the lack of automation support and the scale of real digital object sets.

Consider a set of three pdf files where the meta data known is the identification data consisting of format and format version, as shown in Table 1. Most preservation experts would agree that file 1 and file 2 are similar. It is likely that the preservation risks of file 3 are handled differently. But consider the same three files with additional knowledge provided by deeper characterization, as shown in Table 2. Many experts will consider file 1 and file 3 to be homogeneous and may treat file 2 differently.

The problem is that in a real world scenario, such a collection will be significantly larger, with a complex format profile and many more characteristics. This makes it difficult to comprehend the differences of the objects and divide them into homogeneous sets based on those characteristics, that cause the issues during preservation actions. Only in-depth characterization can provide the necessary information required for effective planning. The goal of such a content profile thus is to provide comprehensive overview of a collection considered for long-term preservation.

2. STATE OF THE ART

Approaches and tools demonstrated thus far are generally restricted to format identification [2] or to small scales of content. Hitchcock et al. argue that although profiles can be based on many different aspects, the one that matters is the file format [4]. While it is one of the most significant properties within a collection, it often does not provide enough information to preserve it successfully. As in the example above, this most importantly applies to content that is homogeneous in terms of the format, where potential failures during the execution of a preservation action come from the subtleties of other characteristics.

Automatic meta data extraction is done by numerous tools,

IPRES 2012 November 2012, Toronto, Canada

Table 1: An example set of three pdf files

Characteristic	File 1	File2	File 3
Format	PDF 1.2	PDF 1.2	PDF 1.4

Table 2: The same set with additional meta data

Characteristic	File 1	File2	File 3
Format	PDF 1.2	PDF 1.2	PDF 1.4
Page count	20	20.000	40
Encryption	Yes	No	Yes
File Size	1 MB	120 MB	2 MB
Valid	No	Yes	No
Well-formed	Yes	Yes	Yes

such as Apache Tika, JHove and many more. The FITS[1] tool follows a different approach that unifies many different characterization tools but provides a normalized output of their results and gives indicators for their validity. These features provide a solid basis for preservation analysis and a complete content profile.

One key argument against the usage of in-depth characterization is that the analysis of meta data produced is extremely time-consuming. This stems from the observation that even the amount of meta data itself may be substantial. However, scalable approaches for content characterization can build on parallel architectures such as map-reduce to increase the processing speed in the analysis itself [3].

3. SCALABLE CONTENT PROFILING

Content profiling consists of three high-level steps: meta-data gathering, processing & aggregation and meta-data analysis. The first step transforms the data in a model that supports faster and scalable analysis and stores it. Post-processing solves issues, such as conflict resolution, due to the normalization of data provided by different tools and aggregation provides a machine readable overview of the data. The last part of profiling offers the planning expert a service on top of the data. It helps the analysis of the subtleties of the objects and partitioning the content into smaller sets fit for a specific preservation action.

Clever, Crafty, Content Profiling of Objects (c3po[2]) is a software tool prototype, which uses FITS generated data of a digital collection as input and generates a profile of the content set in an automatic fashion. It is designed in a way so that different meta data formats originating from other tools can be easily integrated. The tool follows the proposed three part profiling process and provides facilities for data export and further analysis of the content, such as helpful visualizations of the meta data characteristics, partitioning of the collection into homogeneous sets based on any known characteristic. In order to support the decision making it also makes use of different algorithms that choose a small set of sample records (up to 10) based on the size of objects, the distribution of specific characteristics, or other common features. For each chosen partition of the content, a special machine-readable profile can be generated that contains aggregations and distributions for many of the properties. The profile optionally contains the set of chosen representative samples as well as their identifiers within a content repository and a list of all objects that fall into the particular partition. A machine-readable content profile conforming to such a specific format plays an important role for integration with a planning component, content repositories and monitoring systems and thus for the automation of the entire cycle of planning and operations.

c3po makes use of a MongoDB[3] document store, which is a scale-out NoSQL solution. This provides a huge performance boost in comparison to a traditional relational database solution, since the key-value structure of data closely corresponds to the underlying structure of the meta-data collected. The native map-reduce capabilities of this storage solution enable c3po to build format profiles, distributions of any other characteristic and combinations thereof in the order of seconds for hundreds of thousands objects.

To reduce network overhead, c3po offers a command line tool that can be executed near the data and the document store and also a web application, that can be deployed separately and used for visual aid to the planning experts.

4. RESULTS & OUTLOOK

Parsing, post-processing, aggregating, and generating a profile for a medium real world collection consisting of about 42 thousand FITS files (documents) currently takes about 1.5 minutes on a standard PC with 4GB RAM and 2.3 GHz CPU. Similar processing on a much larger real world content set from a web archive, consisting of about 1.3 million FITS files, completes in under 10 minutes on a single machine with 8 CPU cores. Filtering the content based on different characteristics is done via map-reduce jobs, and although it takes 30 to 60 seconds for each job, it turns out to be of great value during analysis.

Future research includes case studies of different content types and algorithms for more effective stratification of samples, integration with Plato[4], as well as with repositories and automated monitoring services. Support and interfaces for different characterization tools and further tool optimizations will provide even more solid basis for faster and scalable analysis.

Acknowledgements

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

5. REFERENCES

[1] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *IJDL*, 10(4):133–157, 2009.

[2] T. Brody, L. Carr, J. Hey, A. Brown, and S. Hitchcock. Pronom-roar: Adding format profiles to a repository registry to inform preservation services. *The International Journal of Digital Curation*, 2(2), November 2007.

[3] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, Jan. 2008.

[4] S. Hitchcock and D. Tarrant. Characterising and preserving digital repositories: File format profiles. *Ariadne*, (66), 2011.

[5] H. Kulovits, A. Rauber, A. Kugler, M. Brantl, T. Beinert, and A. Schoger. From TIFF to JPEG2000? preservation planning at the bavarian state library using a collection of digitized 16th century printings. *D-Lib Magazine*, 15(11/12), 2009.

[1] `http://code.google.com/p/fits/`
[2] `http://github.com/peshkira/c3po`
[3] `http://www.mongodb.org/`
[4] `http://ifs.tuwien.ac.at/dp/plato`