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.

Format profiles are used in the Registry of Open Access Repositories (ROAR), a list of open access repositories of research material. In contrast, it is also possible to purposefully consider the set of properties necessary to describe some aspects of the content.

Decision makers consider data from different aspects, depending on the context and the task at hand. While some may be interested only in volume and number of objects, others are interested in provenance or authenticity. Each aspect requires its own set of properties. For example, Hedstrom et al. [9] introduced 'significant properties' that *"affect their quality, functionality, and look-and-feel so that custodians can select appropriate methods which preserve those significant properties of digital objects that are deemed important by designated user communities"*. The significance of properties may vary in each case, depending on the context and stakeholders [6]. To define which properties are significant, a practitioner must hence possess prior knowledge derived from business goals, policies or planning. C3PO supports a variety of characterization tools and thus enable analysis of different aspects, leaving it up to the decision maker to choose the appropriate set of properties and perspective.

To expand the coverage of properties practically, the straightforward solution is to use several tools that partially characterize the content from different perspectives and provide corresponding metadata. However, combining characterization tools results may be not trivial due to differences in their output schemas, namings, encodings etc. FITS (File Information Tool Set)[1] is an example of the approach. At its core, FITS is a wrapper for other characterization tools such as Apache Tika, DROID, Exiftool, FFIdent, File Utility, Jhove and others. Based on configuration settings, FITS can have different tools run on specific file formats. Extension of tool support for FITS is possible by creating a mapping from a target tool to the FITS XML schema.

When considering data quality, the results from existing tools are far from perfect, and better tools are clearly needed [14]. There is little common understanding of how to test tools in a systematic and rigorous way. A recent experiment in the SCAPE project[2] evaluated several characterization tools [13]. Hutchins [10] describes his activity on testing file characterization tools by comparing results of the tools against each other on the publicly available Govdocs1[3]. The author also raises the issue of lacking standard ground-truth and methods, which made it impossible to check whether a single standalone tool produces correct results. The BenchmarkDP project[4] is developing an approach to generate benchmark datasets for objective, trustworthy validation of properties of characterization tools such as functional correctness [2].

Aggregation and analysis of characterization results is called **content profiling** [12]. Aggregation techniques provide an overview of the content and allow the user to access new knowledge and help explain phenomena surfacing in the

[1] `http://www.fitstool.org`
[2] `http://www.scape-project.eu/`
[3] `http://digitalcorpora.org/corpora/files`
[4] `http://benchmark-dp.org/`

data. Content profiling tools should support the exploration of the content by extraction and analysis of as many characteristics as desired. Rich metadata helps to better describe the content, which may bring additional benefits to preservation processes, such as a more detailed analysis of the content, better requirements specification etc.

C3PO (Clever, Crafty, Content Profiling Tool) [12] is a software tool that enables large-scale content analysis of data collections. Figure 1 describe a general content profiling workflow used in C3PO. The tool uses results from characterization of digital collection as input, aggregates them, generates a profile of a content set in an automated manner. It produces a detailed content profile describing the key properties of the collection.

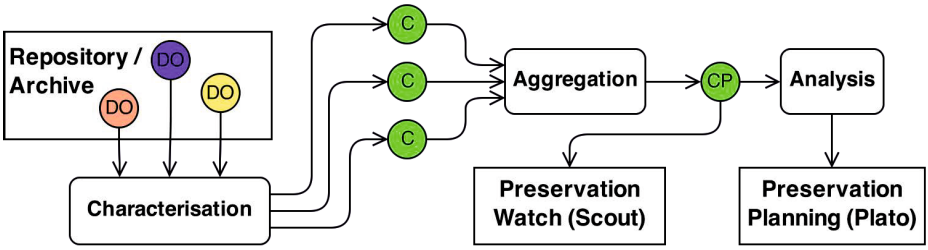

Fig. 1. Content profiling workflow adopted in C3PO

As shown in Figure 1, the workflow starts with running characterization tools on the content. The results are collected and stored by C3PO. Currently, C3PO supports the metadata schema of FITS. C3PO uses MongoDB[5], a scale-out NoSQL solution with sharded cluster calculations and map-reduce support. The content profile generated by C3PO is used in digital preservation tools such as Scout[6] [7] and Plato[7]. Built with an easily extensible architecture, C3PO may be enriched with support for new tools, processing and storing metadata through implementing well-documented APIs. C3PO runs analytical queries to calculate a range of statistics from the size of a collection to distributions of different properties in the collection. The combination of such statistics form a content profile. Basic interactive analytics features are accessible through a web-application. C3PO provides facilities for data export and further analysis of the content, such as helpful visualizations and querying the content characterization results, partitioning the metadata into homogeneous sets based on any captured characteristic selected, and generating representative samples.

The results of content profiling will differ depending on properties chosen for aggregation. For example, we may have a distribution of the MIME-type property values of the collection as in Figure 2. Sometimes it is also necessary to

[5] http://www.mongodb.org/
[6] http://openplanets.github.io/scout/
[7] http://ifs.tuwien.ac.at/dp/plato/

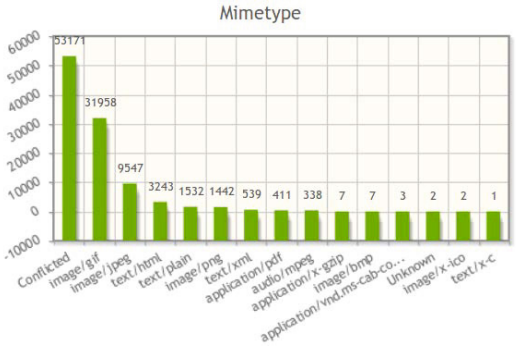

Fig. 2. MIME type property value distribution of the collection

for specific characteristics, for example to classify PDF 1.2 documents according to the applications which created them to identify documents at risk.

A crucial task within content profiling is generating representative samples. Sampling is a process of picking digital objects which represent the whole collection based on certain criteria. For example, sampling may help describe file format name distribution of a collection by picking samples from the 4 most popular format names. Sampling enables controlled experimentation without the need to use the entire collection. Representative samples, the metadata and the digital objects themselves can be used for further experiments without dealing with the collection. This is extremely important in case the collection is of a huge size and you have to run planning process, where preservation workflows should be evaluated on the dataset. Having representative samples, it is possible to test workflows on these samples and have reasonable confidence in their behavior on the entire set without expensive experimentation setup. However, without tool support, criteria have often been based on intuition, prone to individual bias and not based on an understanding of the technical variety of content [3].

Limitations of combining characterization and content profiling emerge from their nature. Most importantly, the overall data quality is dependent on the quality of characterization. If characterization tools do not return correct results, it is not possible for content profiling to provide correct data analysis and insight. However, combining multiple tools should allow us to improve the quality provided the right mechanisms are in place.

3 Challenges and Contribution

While combining metadata from the tools, conflicts will arise in identifying a correct value for a property. This may happen due to several reasons. We group them in 3 overall categories:

No Common Vocabulary. A common problem among tools in the absence of agreed terminology has been the introduction of proprietary vocabulary,

giving new names to concepts, properties and their possible values. For example, characterization tools commonly supply a variety of different labels for the TIFF format. These can all be called 'correct', but pose challenges for further processing.

Specificity of Tools. Some characterization tools may perform on specific content better then others do and provide deeper knowledge. For example, 2 tools report on a file that its MIME type is either "application/xhtml+xml" or "application/xml". Such two results could be deemed completely different, when, in reality, the former is a refinement of the latter.

Conflicting Results. Tools provide competing characterization results, i.e. tool A says a file is a PDF document, while tool B says the file is a TIFF image.

This list is not exhaustive. A deeper classification of reasons of conflicts can be found in [5]. The authors run a case study to analyze the nature of digital object properties that were captured in different preservation institutions.

Apart from these peculiarities, a challenge arises regarding scalable data processing. As the sizes of collections in institutions are increasingly measured in petabytes, traditional methods and database systems are hardly applicable. Doing analytics on such collections becomes more complex, takes more time and requires scalable approaches.

It is also important to note that representative sampling is a challenging task. In order to capture the technical variation in the set of objects, samples should be representative according to more than one dimension. Current tool support for this is scarce, and the quality of input data will limit the sampling accuracy.

To address the given challenges we used and extended the functionality of C3PO by adding the following features:

Rule-Based Engine. As part of the gathering step, once a characterisation result is read by C3PO, the metadata is processed by a plug-in based on the Drools framework before storing in a database. Drools[8] is a business rule management system with a rule engine based on the Rete algorithm[8]. It allows to create an extensible set of human written rules to solve a broad range of business tasks including conflict reduction. Further, this section describes the rules created for conflict reduction.

Vocabulary. Properties stored in C3PO are mapped to the existing vocabulary, PW Ontology[9] [11]. It defines a common list of measures that may help to describe digital preservation context and is used in preservation tools Plato and Scout.

Characterisation Tool Support. Apache Tika[10] was added to the list of supported tools. This allows running fast format identification before doing

[8] `http://www.jboss.org/drools/`
[9] `http://purl.org/dp/quality/measures`
[10] `http://tika.apache.org/`

fully featured and time-consuming characterisation with FITS. Having results from different tools, C3PO is capable to consolidate them to provide more details.

Aggregation Mode. Although MongoDB is a well-recommended database solution for working with large-scale data, it requires some technical background in order to setup a cluster with appropriate data sharding settings. As an alternative we have added a new processing mode, called *DirectProfile* and available in C3PO starting from version 0.5. In this mode, data is processed on the fly without entering a database. C3PO iteratively reads characterisation outputs and accumulates metadata statistics in memory. This allows incremental content profiling with small footprint. Interactive querying and filtering is not applicable in this mode.

When creating a new rule, it must contain 4 elements: name, priority (a number from 0 to 1000), when- and then- clauses. The last 2 elements define correspondingly a list of conditions when a rule should be triggered and actions occurred on a trigger event. Within this work, we have created a list of rules, presented in Table 1. These rules address conflicts in govdocs1 processed by FITS version 0.6.2 and are available in C3PO starting from version 0.5.

Table 1. Identified conflicts per property

Rule ID	Treated property	Target tool	Rule description
1	mimetype, format	Droid	if Droid will report a file format is "Microsoft Powerpoint Presentation", but Exiftool will not report a MIME-type "PPT/S", ignore this identification (Droid alone has false positives on "Microsoft Powerpoint Presentation"). This is a pre-cleaning step to remove wrong mimetypes or formats
2	mimetype	Exiftool, Droid, all	if Exiftool and Droid will both report a file format "Microsoft Powerpoint Presentation", ignore others (format and mimetype), because if the first two tools agree, then the identification is correct
3	format	Exiftool, Droid, all	if Exiftool and Droid will both report a file format "Microsoft Powerpoint Presentation", ignore others (format and mimetype), because if the first two tools agree, then the identification is correct
4	mimetype, format	Jhove, Droid, all	If Jhove and Droid will both report a file mimetype "application/xhmtl", ignore others, because if the first two tools agree, then the identification is correct
5	mimetype, format	Jhove, all	If Jhove will report a file mimetype "text/html" and some other tools will report the file mimetype "application/xhtml+xml", ignore the "text/html" mimetype and the corresponding format
6	format	Jhove, all	If Jhove will report a file format "HTML Transitional" and other tools will claim it to be "Hypertext Markup Language" at least 2 times, "Hypertext Markup Language" is used
7	author	Exiftool, all	If Exiftool will mention file author, ignore others, because Exiftool is correct

4 Experiments

C3PO conflict resolution capabilities were tested on a publicly available data set, govdocs1, which contains approximately 1 million files. For the experiment, we obtain characterization results from running FITS version 0.6.2 on the corpus.

Next, we split the data in 2 parts: first part with 100000 files and second part with 900000 files. The first part is used as a training set, for which conflict resolution rules are created and verified. Testing of the rules is done on the second part of the data. This will shed light on how wide and general the rules can be applied in real world cases.

Firstly, the training set was analysed using C3PO. C3PO reported less amount of processed data, which may happen due to imperfection and bugs in code base of C3PO and FITS. The analysis revealed conflicts in properties identified and reported by FITS. For the experiment, we selected 4 properties with statistics on conflicts, presented in Table 2.

Secondly, a set of rules (see Table 1) was created to address the mentioned conflicts. They were obtained by empirically studying reasons of the conflicts in every case. The most common reasons are addressed in the rules.

Thirdly, we iteratively applied the rules 1-7 from Table 1 to the training set in accumulated fashion: in the first iteration, we applied the rule 1, in the second - the rules 1 and 2 and so on. In total, there are 7 iterations. After each iteration we calculated the amount of conflicts remained after reduction process. The results are presented in Figure 3.

Figure 3 contains a stacked area chart, each area of which corresponds to one of the properties of interest: "Author", "Mimetype", "Format version" and "Format". Statistics are presented in Table 2. Before running experiments, the training set contains 72572 conflicts in characterisation results of 65929 digital objects. After applying the 7 rules, there are 45988 conflicts left, which is 63% of the total amount of conflicts. With respect to the selected properties, 40665 conflicts were reduced down to 11483 conflicts, which is 28% of the initial amount. The chart demonstrates that the rules affect mostly "Format" and "Mimetype" properties, as the amount of conflicts of theirs reduces in every iteration. Conflicts in "Author" property are reduced by the rules 1 and 7. The "Format version" property is affected only by the rule 1.

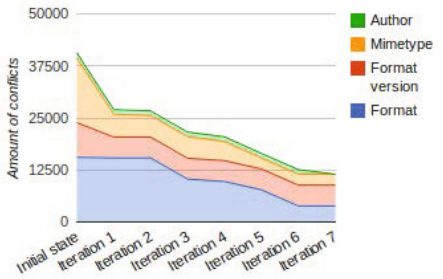

Fig. 3. Amount of conflicts for the given properties in the training set during experimentation

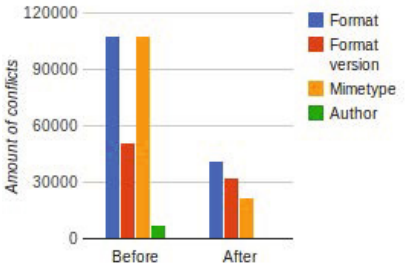

Fig. 4. Amount of conflicts for the given properties in the test set before and after experimentation

Finally, we also want to know how generally the rules 1-7 may be applied on the test set. To check this, we run C3PO with the 7 rules created for the training set on the test set. Figure 4 contains the chart with the results of this experiment. Before running experiments, the test set contains 656412 conflicts in characterisation results of 579587 digital objects. After applying the 7 rules, there are 473022 conflicts left, which is 72% of the total amount of conflicts. With respect to the selected properties, 273723 conflicts were reduced down to 95642 conflicts, which is 35% of the initial amount.

The most effective rule is the rule 7, which resolves 99,8% of conflicts of the "Author" property in the test set. The least amount of conflicts reduced are of the "Format version" property, which is 36% of the total amount of conflict of that property. This is an interesting discovery since there is no single rule that addresses this property directly. The conflicts are mostly covered by the rule 1.

Characterization measures are generally not independent from each other. This can be seen also in the fact that the amount of conflicts in format and mime type is identical: Where multiple tools were able to characterize one file, they generally disagree on how to label it, even if they classify it identically.

From the last experiment we can conclude that the rules created for the training set performed effectively in the test set. It is important to note that this judgment is done based on an expert analysis. The expert studied the content and selected the list of rules that solve certain conflicts. The rules created by experts can be easily shared and assessed in comparative experiments. The heuristics thus enable analysis with much improved data quality.

The test set contains similar proportion of conflicted objects that were identified in the training set: 69% and 70% - before experiments, 40% and 35% - after experiments, correspondingly. The conflicted objects and conflicts are evenly distributed in the FITS characterization results of govdocs1.

Table 2. Conflicts in training and test sets

Set	**Total amount of objects**	**Measurement done wrt experiments**	**Objects with conclicts**	**Amount of conflicts**				
				in a set	in Format	in Format version	in MIME type	in Author
Training Set	96207	Before	65929	72572	15529	8332	15529	1275
		After	34245	45988	3838	5018	2603	24
Test Set	849539	Before	579587	656412	107546	50969	107546	7662
		After	336451	473022	41024	32236	22225	157

5 Summary

In this paper, we discussed and addressed challenges that concern quality, depth and scale of content profiling and presented an approach to improve data quality efficiently by extending the content profiling tool C3PO. We introduced a mechanism for extensible post-processing of metadata based on stateless rule processing engine Drools in C3PO. This engine was adapted to provide conflict reduction capabilities which improves gathered metadata quality. The resulting

rule set is presented during the series of experiments on a publicly available large data set. 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 rule mechanism and the set of rules are part of the publicly accessible c3po, which is freely accessible on github[11].

As a next step, we will evaluate the rule creation mechanism on large real-world datasets. Besides potential further scalability challenges, it is an opportunity to deepen and share the community's knowledge about reasons of characterization conflicts and heuristics to treat them, evaluate how rules from different content collections may improve conflict reduction, and thus contribute to the evidence base of digital preservation. We will also address challenges in representative sample generation, evaluating and selecting appropriate sampling heuristics.

Acknowledgements. Part of this work was supported by the Vienna Science and Technology Fund (WWTF) through the project *BenchmarkDP* (ICT12-046), and by the EU in the 7th Framework Program, IST, through the *SCAPE* project, Contract 270137. The authors would like to thank Petar Petrov for his technical support.

References

1. Abrams, S., Morrissey, S., Cramer, T.: What? So What.: The next-generation JHOVE2 architecture for format-aware characterization. IJDC 4(3) (2009)
2. Becker, C., Duretec, K.: Free benchmark corpora for preservation experiments: using model-driven engineering to generate data sets. In: Proc. JCDL. ACM (2013)
3. Becker, C., Rauber, A.: Preservation decisions: Terms and conditions apply. In: Proc. JCDL. ACM (2011)
4. Brody, T., Carr, L., Hey, J., Brown, A., Hitchcock, S.: PRONOM-ROAR: Adding format profiles to a repository registry to inform preservation services. IJDC 2(2) (2008)
5. Dappert, A.: Deal with conflict, capture the relationship: The case of digital object properties. In: Proc. IPRES, pp. 21–29 (2010)
6. Dappert, A., Farquhar, A.: Significance is in the eye of the stakeholder. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (eds.) ECDL 2009. LNCS, vol. 5714, pp. 297–308. Springer, Heidelberg (2009)
7. Faria, L., Petrov, P., Duretec, K., Becker, C., Ferreira, M., Ramalho, J.: Design and architecture of a novel preservation watch system. In: Chen, H.-H., Chowdhury, G. (eds.) ICADL 2012. LNCS, vol. 7634, pp. 168–178. Springer, Heidelberg (2012)
8. Forgy, C.L.: Rete: A fast algorithm for the many pattern/many object pattern match problem. Artificial Intelligence 19(1), 17–37 (1982)
9. Hedstrom, M., Lee, C.A.: Significant properties of digital objects: definitions, applications, implications. In: DLM-Forum, vol. 200, pp. 218–27 (2002)
10. Hutchins, M.: Testing software tools of potential interest for digital preservation activities at the national library of australia. NLA Australia Staff Papers (2012)

[11] `https://github.com/openplanets/c3po`

11. Kulovits, H., Kraxner, M., Plangg, M., Becker, C., Bechhofer, S.: Open preservation data: Controlled vocabularies and ontologies for preservation ecosystems. In: Proc. IPRES, pp. 63–72
12. Petrov, P., Becker, C.: Large-scale content profiling for preservation analysis. In: 9th International Conference on Preservation of Digital Objects (IPRES 2012) (2012)
13. van der Knijff, J., Wilson, C.: Evaluation of characterisation tools. part 1: Identification. Technical report, National Library of the Netherlands (2011)
14. Wheatley, P.: The practitioners have spoken: "we need better characterisation!". Blog post (2012), `http://www.openplanetsfoundation.org/blogs/2012-10-19-practitioners-have-spoken-we-need-better-characterisation` (accessed June 2014)