Dagstuhl Seminar 10291: Automation in Digital Preservation, 18-23 July, 2010

Digital Objects Characterization: Document Conversion and Quality Assurance

Natasa Milic-Frayling
Microsoft Research Cambridge
Roger Needham Building, 7 J J Thomson Avenue, Cambridge CB3 0FB, United Kingdom
natasamf@microsoft.com

Abstract. Whether we are migrating document formats to achieve interoperability or ensure long term preservation, we are faced with the issue of assessing the quality of the digital object transformation. However, comparing two digital objects is not straightforward. It raises the issue of properties that are inherent to the digital objects and those that are dependent on the environment in which the objects are created, viewed, and compared to one another. That has implications for devising methods to extract document properties, interpret observed characteristics, and apply similarity metrics. Furthermore, in order to take actions based on collected measurements, we need to define or learn the significance of individual document properties from the perspective of human perception and usage scenarios. We illustrate the complexity of these issues by presenting a method for comparing converted office documents and discussing the challenges from the technical and methodology point of view.

Introduction

With a broad proliferation of digital media applications and services, the variety of digital objects continuously increases, encompassing data, media content, scripts, programs, and similar. However, they all share one thing in common. They need to be persisted during processing and then stored for subsequent use. Their instantiation is possible within computing environments that can read the data store and interpret the file format. The issue arises when the supporting computing environment changes or the object needs to be used within a different environment.

For example, a drawing in jpg format, produced by the Paint software on a desktop, can be viewed on a mobile device using a picture viewer but cannot be edited unless there is a compatible drawing application that can process .jpg files. Thus, it is the absence of compatible software that prevents certain usage of the content. Similarly, if data stored in MS Excel spreadsheet needs to be published as a Web page, it first needs to be converted into HTML format that Internet browser can import and render.

The problem of the object re-instantiation is exacerbated when the document store is disconnected from the evolving computing environment over an extended period of time. That is the case in long-term preservation scenarios. Ensuring that a digital object can be accessed and used in the far future requires that we devise software that can convert the original file format into one that is supported in the contemporary environment. That is referred to as *document format migration*. Alternatively, we

need to maintain an environment in which the original software can be instantiated and used, e.g., through *software emulation*. In the latter case, the object is consumed in its original form but confined to the legacy environment. On the other hand, transforming it to a contemporary format is likely to involve a loss of original characteristics.

Characterisation of digital objects is concerned with measuring properties that are significant for the utility of the digital object ([4][6][7]). While this notion is promoted in the context of long term preservation, many of its aspects apply to other scenarios, such as content interoperability, and these will be of our interest. In order to measure a possible loss of fidelity during document format transformation, we need to define relevant properties and develop effective tools to measure attributes of such properties. Experts in preservation planning have developed frameworks for grouping document object characteristics and specifying the weights that reflect the importance of the properties. However, there are very few tools available for measuring specific properties and computing the corresponding metrics.

In the following section we describe a method for comparing visual properties of office documents and outline several key issues associated with measuring them: (1) defining a common representation to serve as the basis for characterization and (2) understanding the potential loss of information caused by intermediary mappings during object representation, feature extraction, and metric computation.

Document Conversion and Common Representation

Open source community has created a number of document conversion tools that can be used to convert document from one format to another. Most recently, they have built converters from the proprietary format of WordPerfect and MS Word to XML based formats such as OpenXML and Open Document Format (ODF). The converters are written with a goal to preserve various aspects of the original document format.

Through our engagement with the PLANETS project, we adopted the notion of document object *aspect* to designate an abstract view of the information, referring to the content, structure, interaction, and similar [4]. For example, the spacing between characters in a WordPerfect document or metadata of MP3 files are examples of *explicit aspects* that are, in fact, parts of the files and can be detected by inspection. However, it is often useful to compute *implicit*, i.e., *derived aspects* using processing of the data in the objects. Such are the character count in Word documents, colour histograms for bitmap images, and similar. Comparison of explicit aspects is more straightforward because they are likely to be in the same place and same form in both files. In some instances, specific aspect may be universally present with all files, such as last modified time.

Document Conversion Service. In order to facilitate documents conversation, we created a service-oriented architecture that exposes the document conversion and comparison as Web services. This Web portal enables individuals to submit documents for conversion and to view the original and the converted digital objects through the Web interface. More precisely, the Web viewer exposes images of these objects that are rendered by their native applications on the server side. Then for a specific aspect, related to the visual characteristics of the document, we analyze the images and highlight the differences.

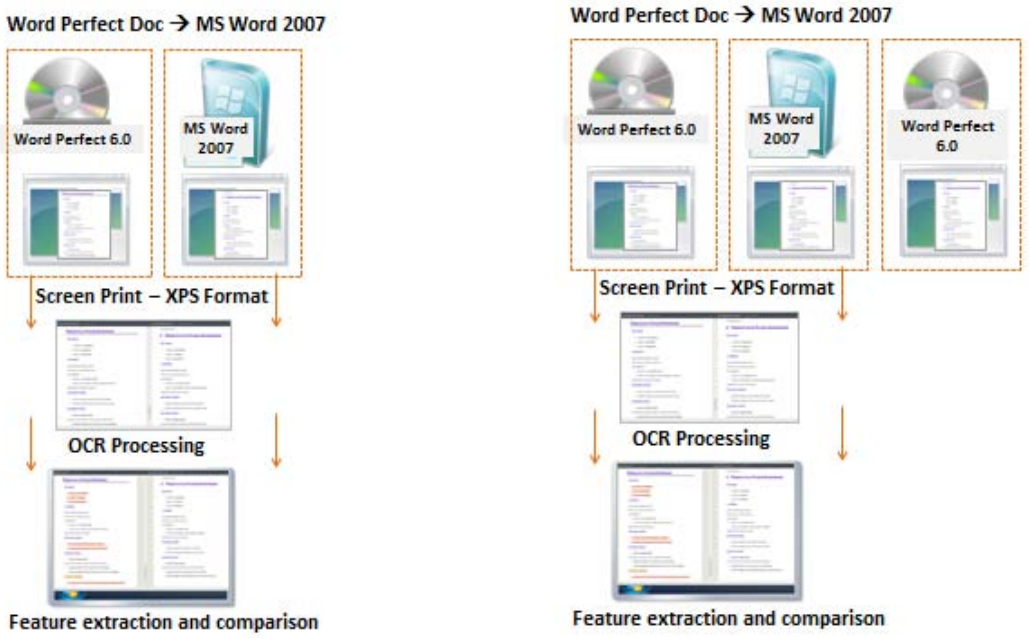

Figure 1. (a) Two objects in different formats are mapped onto a normalized form. For each feature we develop a `digital object probe' that extracts the feature and measures a property of the feature. That is facilitated by OCR software (b) Output of the emulator can be added and compared using the same technique: image of the emulated randering can be passed through the OCR and feature extractor.

Characterizing a specific aspect of the digital object can often be achieved by repurposing tools that already exist and are used to process the files. Most useful are the tools that can run within both environments, of the original and the converted document, and produce the same or similar type of information about the given aspect. In our example we focussed on characterizing the layout features and used the commercial, off the shelf software for Optical Character Recognition (OCR) software to process the screen-capture images of both documents.

Comparison of Digital Objects

Figure 1 illustrates the workflow for comparing the layout aspects of the original document A, in the Word Perfect format, and the converted document B, in the Word 97 format. We first created a common representation of both digital objects as an image format. This representation seemed most appropriate as it preserves well the visual characteristics of the document layout.

The process involves several steps:

- In the environment *E1*
  - We instantiate the document F1 in its native application A1
  - We create an image I1 of the document F1, rendered in the application A1.

- In the environment *E2*
  - We run the conversion of the document F1 to the document F2.
  - We instantiate F2 in the application A2 and create an image representation I2.
  - We compare the two images I1 and I2 by applying a comparison tool.

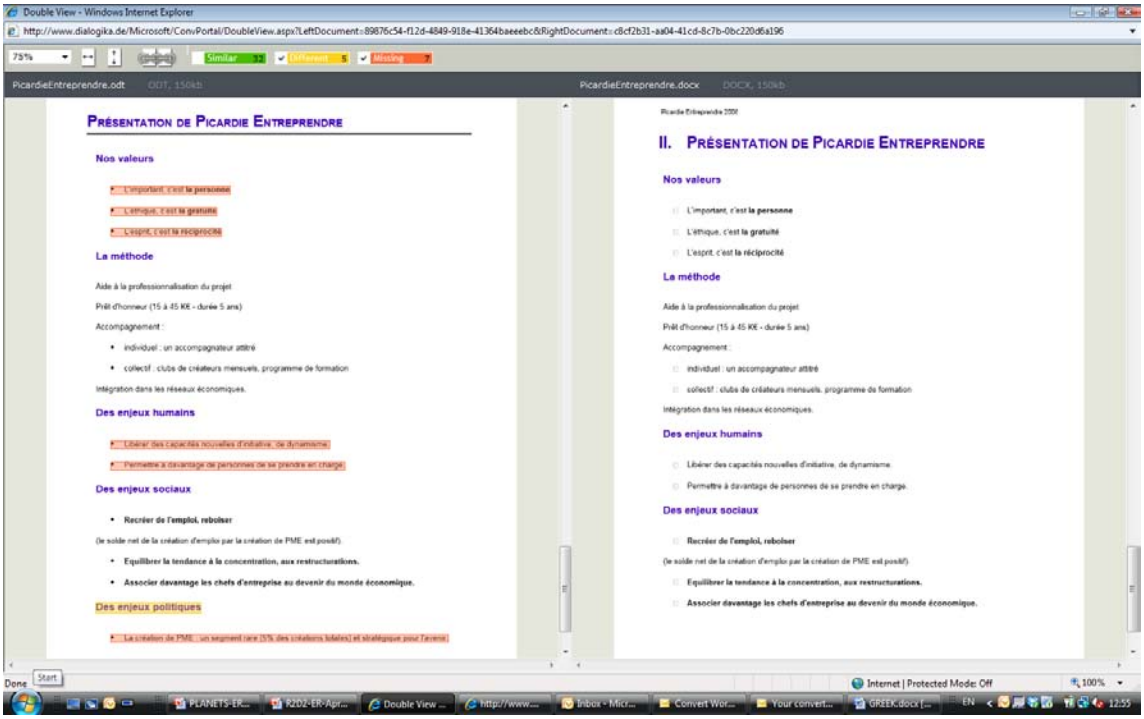

Figure 2. A comparison of the layout features, facilitated by the analysis of the OCR outputs for the images of original and converted document.

The intermediary image representations were captured as XPS formats (I1 and I2) of the 'printing' files (F1 and F2) and analysed with the OmniPage OCR package. We use OmniPage recognition of text blocks as the basis for the content comparison and identify sections of text that differ in two documents. This information is presented as a layer on top of the XPS images (Figure 2).

The rich output of the OCR process supports a number of other content analyses. The layout characterization is particularly interesting because it highlights a couple of specific issues.

- The precision of the characterization method depends on the accuracy of OCR process and our algorithms for content comparison.
- Definition of effective characterization metrics is difficult. One can provide a simple aggregation metrics to indicate a level of discrepancy or a distribution of errors across the document. However, the most comprehensive and practically effective way to present the comparison is by highlighting regions of dissimilarity (Figure 2). Yet, drawing regions onto the image is another source of potential error.
- A detailed review and assessment of layout differences can be accomplished by humans through visual inspection. However, this is prone to error and for long documents may not be feasible or economical.

Measurements and Methodology. In practice, a comparison of two digital objects is performed through a serious of independently applied measurements [4]. As the two objects reside in two different computing environments, it is possible to compare them if:

- In each environment, there is a tool that can measure the same property of the object and provide output useful to compare the two, or

- We can map the two digital objects onto a common representation and apply a single tool to measure and compare a given property of both files.

Generally, once a common representation exists across environments, one can use the same *technical probe* to extract properties of each object and compare them according to a specified metric. One should keep in mind that using an intermediary representation enables only indirect measurements of the object properties and, therefore, may not be helpful to infer related characteristics of the individual objects. For example, converting a Word document into an image does not let us detect a tooltip that appears on hover and displays further information about an entity. However, we can use the image properties to detect discrepancies between two objects. Thus, while the quality of individual digital objects in an absolute sense cannot be fully characterized, selected types of error, i.e., *difference between the original and the converted document* can be detected.

The advantage of this approach is that, as long as each computing environment can produce the same intermediary representation of the object, we can use a single tool to compare documents from different environments. In Figure 1(b) we show that besides the original and the converted file, one can produce a screen rendering of the same document viewed in the emulated application and therefore evaluate not only the format conversion accuracy but also the emulation approach for that document type.

Should we want to use the measurements from the intermediary representations and learn about properties of the original objects, we need to understand the nature of the mappings that are involved in the process and consider the inverse function for the collected measurements. In our method we use

- The mapping of the rendered object onto the image representation
- The mapping of the image onto the OCR characterization
- The function that extracts or derives a specific property from the OCR output, e.g., the position of the sentences on the page.

At every stage there is a level of complexity, uncertainty, and a loss of information relative to the starting point.

Binding of Document Object Aspects and Software. In practice, the same document object can be viewed using multiple viewing applications. Therefore, document rendering may differ from one viewer to another. When comparing two document objects we may use multiple viewers and therefore obtain multiple data points. Unless there is a clear justification for referring to one of the document views as authentic, we are dealing with a multi-value characterization of the document similarity, each relative to a selected viewer. That brings us to a more general observation:

A given aspect of the data object may be inseparable from the software that instantiates the object, such as a document viewer or a content player. The comparison method and the measurements are relative to the specific viewer. Using multiple viewers presents alternative representations and therefore opportunities to investigate further properties of the digital objects.

Discussion

Economics of content transformation, characterization, and quality control, is subject to the cost-benefit trade-offs and raises an important question: how important is it to have a complete and high quality characterization of a digital object? The answer is likely to depend on the context and the usage scenario. Thus, in addition to the technical issues, we need a deeper understanding of the

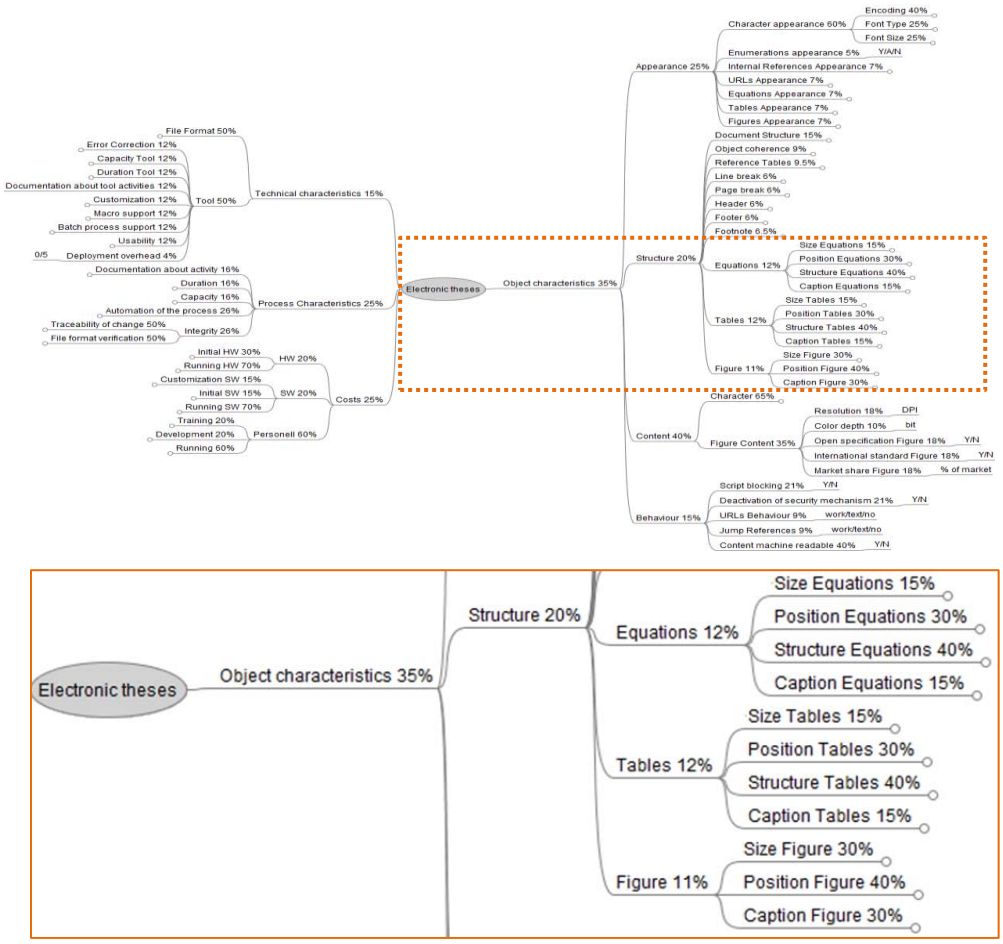

Figure 3. Taxonomy of significant properties derived for the Electronic Theses digital objects (Becker et al. [3])

importance that longevity of digital objects has for the human kind and which aspects are essential and need to be preserved.

Resorting to the case studies and analyses of digital object documentations, experts are mapping out taxonomies of significant properties (Figure 3, [3],[8]). For each individual property one has to devise a tool to collect measurements, define metrics to enable characterization or comparison, and specify the relative significance of that property within the overall metric calculation. Yet, to enable characterization and comparison of any digital object, these properties need to be expressed in a format-neutral way. Thus, the PLANETS project included a definition of an *extendible characterisation definition language* (XCDL) that allows the practitioners to grow the schema of significant properties and the *extensible characterisation extraction language* (XCEL) that enables the practitioners to implement extractor components that can interpret this language.

Becker et al. ([1][2]) describe the principles of using XCDL and XCEL. In effect, for a given object in the file format F, an XCDL document provides an XML description of the file's content according to the XCDL language specifications. That document can be processed using an XCDL interpreter. On the other hand, the XCEL document describes what information can be extracted from any file of format F. Thus, it is used by an XCEL processor to extract this information and express it in XCDL. In other word, the XCEL creates a mapping between the declarative description of the information in a physical file and its abstract interpretation outside of a format specification.

Digital Interpreters and Human Perception. Building on the PLANETS characterization framework we developed several extractors of document features and realized that the meaning and significance of the extracted measurements need to be determined by the user. That introduces another level of complexity. Frist, interpretation of digital content by a human is possible only through intermediary software, e.g., content viewer for text, content player for audio and video, and haptic surfaces including touch based screens and devices. Indeed, the humans can consume digital information through a few key senses and their observation of digital content is subject to the human perception facilitated through these senses. This leads to the issue of mapping a programmatic characterization of the file content (e.g., simple comparison of text boxes) onto the perceptive characterization by a human observer.

In contrast to the repeatable and deterministic comparison of two objects facilitated by software, even a single document characteristic, e.g., a font colour, can be perceived differently across the human population. This is not necessarily an issue in a comparison task when the same individual is viewing two documents in identical conditions. However, it is important to exploit and reconcile the differences in the human perception in order to arrive at the characterization of digital object that can be useful across users.

Crowdsourcing and Community Relevance. For some aspects of digital objects it is valuable to collect and aggregate views across individual of a particular user segment or a random sample of users. For example, preserving the document object model associated with a file format may not be critical for readers of the document content. However, that may be absolutely essential for individuals studying how digital object representation varied over time across software applications. This calls for a crowdsourcing approach across communities of practice or other segments of the population.

Automation. In order to achieve the scale, we need to develop automated methods for quality assessment that leverage human input on a specific set of digital objects and extrapolate to the unseen ones. This naturally leads to the combination of pattern analysis and machine learning techniques that can become instrumental in arriving at economically viable approaches to digital object conversion.

Based on our initial research investigations, we are led to interesting and challenging research questions:

- What is the relationship between the human criteria and automated measurements of content similarity?
- What influence do usage scenarios have onto the approaches to digital object characterization? Can we define a set of fundamental and object inherent properties that can be used to derive high level characteristics?
- What are the properties of ‘instruments’, i.e., technical probes that we need in order to extract and measure properties of digital content? If the characteristics are measure indirectly, what are the properties of the intermediate mapping and compatibility criteria with the metrics used to characterize them?
- How can we extend our preservation objectives to include characterization of digital objects that are volatile in their nature, such as content streams, or complex, such as networks resulting from the social interactions in online computing environments?

References

1. Becker, Ch., Rauber, A., Heydegger, V., Schnasse, J., and Thaller, M. A Generic XML Language for Characterising Objects to Support Digital Preservation. SAC'08, March 16-20, 2008, Fortaleza, Ceara, Brazil.
2. Becker, Ch., Rauber, A., Heydegger, V., Schnasse, J., and Thaller, M. Systematic Characterisation of Objects in Digital preservation: The eXtensible Characterisaction Languages. Journal of Universal Computer Science, vol. 14, no. 18 (2008), 2936-2952.
3. Becker, Ch., Strodl, s., Neumayer, R., Rauber, A., Nicchiarelli, E., and Kaiser, M. Long-term Preservation of Electronic Theses and Dissertations: A Case Study in Preservation Planning. (2007) The Ninth Russian National Research Conference, RCDL'07, October 15-18 2007 in Pereslawl, Russia.
4. Brown, A. Developing Practical Approaches to Active Preservation. The International Journal of Digital Curation, (2007) (1), vol. 2.
5. Clausen, L. R. Opening Schrödingers Library: Semi-automatic QA Reduces Uncertainty in Object Transformation L. Kovács, N. Fuhr, and C. Meghini (Eds.): ECDL 2007, LNCS 4675, pp. 186–197, 2007.
6. Farquhar, A. and Hockx-Yu, H. Planets: Integrated Services for Digital Preservation. The International Journal of Digital Curation, (2007) 2, vol. 2, 88-99.
7. PLANETS Project (2006) http://www.planets-project.eu.
8. Strodl, S., Becker, C., Neumayer, R., and Rauber, A. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL'07) (June 2007), pp. 29–38.

Dagstuhl Seminar 10291

Automation in Digital Preservation

18-22 July, 2010

Organizers:
Jean-Pierre Chanod, Xerox Research Center Europe - Grenoble, France
Milena Dobreva, University of Strathclyde - Glasgow, Great Britain
Andreas Rauber, TU Wien, Austria
Seamus Ross, University of Toronto, Canada.