# PRESERVING VISUAL APPEARANCE OF E-GOVERNMENT WEB FORMS USING METADATA DRIVEN IMITATION

**Jörgen Nilsson**
Luleå University of Technology
Dept. of Business Administration and Social Sciences

## ABSTRACT

This paper summarizes work done in a PhD study on metadata driven imitation for preservation of visual appearance of web forms and/or receipts used in eGovernment services. The research done suggests that metadata, and e.g. a background image, can be used to describe the visual appearance of documents, and that this also facilitates having the data separated from the visual appearance. This separation provides the ability to present the material to the users in different ways, depending on their needs and requirements, while retaining the ability to present the object in its original look. The original look is seen as the most versatile way of presenting the material, giving the most fruitful base of interpretation and understanding, but if the users where familiar with the material, they liked the ability to have the material presented in simplified ways, where many of the sometimes "distracting" visual attributes where removed. In general, preserving the visual appearance *and* keeping the data separated from the form, was seen as useful and beneficial to both the users and the preservation professionals. As always in digital preservation contexts, documentation of this process and the relation between the metadata describing the visual appearance and the data of the document, is of high importance.

## 1. INTRODUCTION

In recent years, the ongoing eGovernment proliferation of public administration has taken great steps toward availability and sophistication. The eGovernment Benchmark Survey 2009 [4] shows that the overall level of full online availability of 20 basic services in the EU27+ has risen from 59% to 71% between 2007 and 2009. The sophistication of the services has risen from 76% to 83% in the same period of time [4]. These are average numbers for the EU27+, some countries have achieved 100%, and yet some are over 90% in both categories.

This increase reasonably means that there will be an increase in the number of digitally born documents/records in need of preservation. Some organisations might also need/want/be obligated to preserve the visual appearance of these documents, and maybe also the appearance of the services, in order to fulfil expectations and demands from their designated community. There can be numerous reasons to preserve the visual appearance of digitally born documents, and some of them can be found in reasoning around the concept of *information*.



The concept of information has in this work been influenced by the *infological equation* (1) which states that *information* (I) is the result of an interpretation process (i) that acts upon data (D) involving the parameters of pre-knowledge (S) and time (t) [12].

$$I = i(D, S, t) \quad (1)$$

One important implication of the infological equation is that data does not contain information but at best can represent information to those who have the required pre-knowledge [12]. In addition to this data also acts as *constraining affordances* where data allows some constructs of information and impede others and that these constructions might differ between individuals [5]. Since humans interpret data, and occasionally with different results, as much of the original data should be available in order to give good basis for similar interpretations by different individuals. Part of this original data can exits in the form of visual attributes, such as colour, italics, tables and other layout properties.

This has lead to an interest in preserving "looks" of web resources, especially those created in eGovernment services.

## 2. PRESERVATION OF WEB

There are (at least) two approaches to web preservation. One approach consists of gathering the web-site(s) with a crawler accessing the web as a client and thereby fetching the web from a user perspective by following links. A typical drawback with crawling would be that it does not fetch documents that you as a user would need to fill in a form to fetch (i.e. deep web), for example by searching in an article database [9].

Another approach to gathering the web would be to keep the server side of the web intact, meaning that the web site still could be accessible in it's original way, as long as the ability to run the entire server side, including e.g. databases, still exists [9]. This could be facilitated

by the use of emulation or migration depending on the requirements of the organisation. The emulation approach has for quite some time been proposed as *the solution* [15], but as pointed out both migration and emulation is not yet mature enough for large scale preservation scenarios, although it usually is better than doing nothing [9].

### 2.1. Significant properties

This paper assumes that the visual appearance or physical structure of a digital object (e.g. a web form) is considered to be a significant property of the object. This may of course differ from case to case as with all significant properties [2], and is certainly not true in all preservation of digital objects. Significant properties are "those components of a digital object deemed necessary for its long-term preservation" [2]. This is a quite common view of significant properties [6],[11],[16], held on a generic level since it is hard to be specific about significant properties in writing unless you actually consider one particular object or group of objects.

One way to handle significant properties have been addressed in work with the Underlying Abstract Form (UAF) [8]. The UAF holds "all the significant properties of the data, and is independent of the medium upon which the data is written" [8], and although not mentioning metadata or physical structure explicitly, they do suggest utilizing the *representation information* container in the OAIS model to hold the UAF, which implies using metadata, even though it could be as simple as referring to a viewer application for the data object e.g. Acrobat Reader for a pdf-file. The UAF prefers to have the representation information pointing out the original software used to access the data object, and that this software also should have been preserved. And although "enabling meaningful access to the preserved object includes such processes as recreating the experience of viewing (or even interacting with) the original" [8], the author of this paper however prefer to focus on the viewing part, using an abstraction of the original objects presentation, described with the aid of metadata and e.g. screen dumps, since the original software could mean that you, for good or bad, preserve the system instead of the information [1], meaning that the users in the future would need to know how to use old software in order to access the information. The approach suggested below instead allows for several different ways of presenting the material to the user, depending on their needs and wants.

### 2.2. Preserving physical structure of deep web documents

Although deep web can contain lot of different types of digital objects, a respectable amount of the objects created in eGovernment context would likely be of a textual character related to filling out web based forms. Some of the objects may be e.g. pdf-files submitted as attachments to a web-form, but still – the actual web form would also have some content filled in and saved, most likely, in a database. This implies that we already here have a separation of the physical structure and the data, and when they are combined together again we get the digital object in its original shape [14] or performance [7].

The separation of physical structure and data makes it possible to treat the respective components according to their preservation needs. However, if the intention is that the original shape of the object should be possible to present again to the designated community, you do need to retain the ability to combine them together again in the future, regardless of what preservation actions they have been subjected to.

One way of addressing this re-presentation is to use *metadata driven imitation* [13] where the physical structure is described by a combination of layout metadata and e.g. backdrop images making out the main part of the layout. One could argue that this poses problems regarding the integrity of the document, but as pointed out in the InterPARES project, “a record has integrity when it is complete and uncorrupted in all its essential respects” [10] meaning that the record does not need to be exactly the same as when it was created, as long as the message it communicates remains is unaltered [10].

The type of metadata driven imitation that is mentioned here is most suitable for documents that appear in large numbers with similar physical structure, in other words, typical forms filled out in eGovernment contexts. Bearing in mind that these types of objects usually are not available to web crawling, these deep web objects need to be collected in some other way.

By describing the layout with metadata and background images, the data can then be linked (again, with metadata) to the layout in order to be presented upon request as a "whole". This also facilitates making other sorts of presentations to fulfill requests from different user communities, where some may only need e.g. a particular data field from thousands of forms, while others are more interested in a complete form with its visual appearance as intact as possible. These kinds of diverse user communities are likely customers of e.g. large national institutions such as national archives or national libraries where the *general public* is the *designated community*.

## 3. OPINIONS ON METADATA DRIVEN IMITATION

Studies done on what potential users and preservation professionals think about the approach with metadata driven imitation [14] shows some interesting results that are presented below.

Most users preferred to have the data presented in a simplified form, where some visual attributes where removed (e.g. background colours and logotypes) while the layout in general (i.e. the physical relation between the data elements) remained intact. It should however be

noted that the respondents said that the original look would give the best possibilities for interpretation, depending on the users familiarity with the material. The preservation professionals did prefer the original look, for the same reason as the users; it provides the best basis for a "correct" interpretation. This can be put in relation to the constraining affordances of data, which both facilitates and limits the interpretations possible.

Both the professionals and the users liked the ability to present the material in different ways, depending on the needs of the user. Some would for example only need the data, and cared less for the look of the document for their own purposes, but they also recognized the importance of retaining the ability to represent the document in its original form. The flexibility in presentation is facilitated by the separation of data from physical structure, and as pointed out by the preservation professionals, this separation also facilitates the ability to handle the data and the physical structure in different ways from a preservation perspective.

The separation mentioned above was recognized as a good feature from a slightly different perspective as well. The ability to only fetch data from a document, mean that it is quite easy to request the same kind of data from a large number of documents, for e.g. statistical purposes, instead of having to extract the data from an actual document, perhaps in an entirely manual way (i.e. actually *reading* the documents). So, although original look was regarded as important in general, the ability to choose from several different ways of presenting the data was seen as valuable. The objects used as demonstrators did not represent the *feel* of the documents, and the users did not see feel as that important on document level, though it certainly can be important at a system level, if that is what you are preserving.

Questions were also posed about the relation between original look and trust. Though the users said that the most trustworthy representation was the original look, they also realized that this might be a false sense of trust. They also pointed at that the trust mainly lies in that they trust the organization that manages the objects, and that they thereby probably would not question a document coming from them that much, in case they did not actually see something that they know is wrong. The preservation professionals, and some of the users, where careful to point out that you must have documentation about the processes concerning the material, for example about how the metadata descriptions of visual attributes are constructed, and used, so that the knowledge about this does not disappear over time.

To sum it up it;

- keep data and physical structure separated for usefulness and flexibility
- find a "middle way" of representing physical structure of the document type in question (e.g. by using a background image for capturing some of the physical structure)
- document everything that the object is subjected to

It can in general be summed up as, yes visual appearance of web forms in eGovernment context is important to preserve, since it both provides more context and acts as constraining affordances and thereby may facilitate better interpretation of the data into the intended information. However, fixing the data to a physical structure may impair the ability to mass process it, and therefore a separation of data from its physical structure would be beneficial. One way of addressing these issues can be by using *metadata driven imitation*.

## 4. REFERENCES


[1] Bearman, D. "Reality and chimeras in the preservation of electronic records." *D-Lib Magazine 5(4).* 1999. Retrieved 2004-03-13 from http://www.dlib.org/dlib/april99/bearman/04bearman.html

[2] Cedars Project. *Cedars Guide to Digital Collection Management*. 2002. Retrieved 2008-04-08 from http://www.leeds.co.uk/cedars/guideto/collmanagement/guidetocolman.pdf

[3] Dollar, C.M. *Authentic electronic records: Strategies for long-term access*. Cohasset Associates Inc., Chicago, IL, 2000.

[4] European Commission. *Smarter, Faster, Better eGovernment*, Brussels, Belgium, 2009.

[5] Floridi, L. "Semantic Conceptions of Information". *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.). 2008. http://plato.stanford.edu/archives/fall2008/entries/information-semantic/

[6] Hedstrom, M. & Lee, C. "Significant properties of digital objects: definitions, applications, implications", *Proceedings of the DLM-forum 2002 Access and preservation of electronic information: Best practices*. pp. 218-223. European Communities. 2002. Retreived 2006-05-28 from http://ec.europa.eu/comm/secretariat_general/edoc_management/dlm_forum/doc/dlm-proceed2002.pdf

[7] Heslop, H. Davis S. Wilson, A. *An Approach to the Preservation of Digital Records*, National Archives of Australia, Canberra. 2002 Retrieved 2004-03-03 from http://www.naa.gov.au/recordkeeping/er/digital_preservation/Green_Paper.pdf

[8] Holdsworth, D. & Sergeant, D.M. *A Blueprint for Representation Information in the OAIS Model*, The Cedars Project, 2000. Retrieved 2004-03-15 from http://esdis-it.gsfc.nasa.gov/MSST/conf2000/PAPERS/D02PA.PDF

[9] International Internet Preservation Consortium. *Long-term Preservation of Web Archives – Experimenting with Emulation and Migration Technologies*. International Internet Preservation Consortium, 2009. Retrieved 2010-04-25 from http://netpreserve.org/publications/NLA_2009_IIPC_Report.pdf

[10] InterPARES. *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES project. Appendix 2 – page 2.* InterPARES Project, 2002. Retrieved 2007-05-10 from http://interpares.org/book/interpares_book_k_app02.pdf

[11] Knight, G. *Framework for the definition of significant properties.* InSPECT project. 2008. Retrieved 2008-04-08 from http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf

[12] Langefors, B. *Essays on infology: summing up and planning for the future.* Studentlitteratur, Lund, 1995.

[13] Nilsson, J. & Hägerfors, A. "Metadata Driven Presentation of Digital Documents/Records", *Constructing and Sharing Memory: Community Informatics, Identity and Empowerment*. Stillman, L. & Johanson, G. (ed.), Cambridge Scholars Publishing, 2007.

[14] Nilsson, J. *Preserving Useful Digital Objects for the Future*. Luleå University of Technology, Luleå, Sweden, 2008

[15] Rothenberg, J. & Bikson, T. *Carrying Authentic, Understandable, and Usable Digital Records Through Time.* The Dutch National Archives and Ministry of Interior, 1999. Retrieved 2008-04-27 from http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf

[16] Wilson, A. *Significant properties report.* InSPECT project. 2007. Retreived 2008-04-08 from http://www.significantproperties.org.uk/documents/wp22_significant_properties.pdf