How open are local government documents in Sweden? A case for open standards.

Björn Lundell and Brian Lings
University of Skövde, Sweden
{ bjorn.lundell, brian.lings }@his.se,
WWW home page: http://www.his.se

Abstract. There is in Europe an increasing recognition of the need for governmental organisations to support and promote the effective curation of electronic data, including public documents, for easy public access and reuse. Such a vision can stand in stark contrast with reality. In this paper we address the question: to what extent are local government documents preserved electronically for discovery and re-use? Our goal is to establish the level to which calls for the greater use of open document standards is being heeded, and to understand the potential consequences of not heeding the advice. We find that availability of electronic copies of documents is very variable, and accessibility is poor. In particular, there is little evidence of policy to maintain electronic copies of documents, and little awareness of open standards and their importance in data curation. This is in stark contrast to stated central Government policy. The study highlights a lack of strategy in organisations regarding the effective curation of electronic data.

1 Introduction

On 26th March 2009 the Swedish Government took a decision to set up a delegation for e-Governance (Regeringen 2009). The delegation was mandated to first draft a strategy for e-Government. In doing so, a number of principles were to be observed. Notable amongst these were principles relating to:

- accessibility and usability in ICT;
- the use of open standards and "software based on open source software";
- solutions that gradually liberate the administration from dependence on specific platforms and solutions;
- long-term digital preservation.

The proposed strategy was published 19th October 2009 as SOU 2009:86 (SOU 2009) and contains strategies for considering open source and open standards in public sector procurement. The reference to open standards is important; standardisation is not in itself considered to be sufficient. The definition of an open standard in SOU 2009:86 (SOU 2009) is identical to that included in the European Interoperability Framework version 1.0 (EU 2004), namely:

1. The standard is adopted and will be maintained by a not-for-profit organisation, and its ongoing development occurs on the basis of an open decision-making procedure available to all interested parties (consensus or majority decision etc.);
2. The standard has been published and the standard specification document is available either freely or at a nominal charge. It must be permissible to all to copy, distribute and use it for no fee or at a nominal fee;
3. Intellectual property – i.e. patents possibly present – of (parts of) the standard is irrevocably made available on a royalty-free basis;
4. There are no constraints on the re-use of the standard."

The principles and strategy are in line with current best practice on the transmission and archiving of electronic data (see, for example, DCC (2005)). The development of effective strategies is seen as essential for the preservation of access to data long-term, something which is of particular importance in the public sector.

Sweden is not alone in setting such objectives. Belgium, the Netherlands, Denmark and Norway already have guidelines on the use of an ISO approved open format for documents used in public administration. PDF and ODF have been specifically identified amongst the few applicable formats. According to Morten Andreas Meyer, Norwegian Minister of Modernisation, in a press announcement on 2nd July 2009:

> "When exchanging documents attached to emails between the Government and users, it is from 1 January 2011 mandatory to use the document formats PDF or ODF." (Regjeringen 2009)

Many articles have been written about the problem of legacy data, i.e. data for which the originating software or hardware is no longer available. Such data is at best difficult and costly to recover, and at worst no longer accessible. In practice, many organisations recognise the potential problems this may cause. In the words of Gordon Frazer, managing director of Microsoft UK:

> "Unless more work is done to ensure legacy file formats can be read and edited in the future, we face a digital dark hole." (BBC 2007)

The need is pressing in the case of open standards for document formats, not just for public access to current data but also for maintaining the archives of data received and generated by governmental organisations.

In this paper, we consider the extent to which, in the absence of a clear policy, electronic access to local government data is being supported. We do this through a quantitative and qualitative study of the availability and accessibility of electronic copies of executive board meetings of Swedish municipalities.

2 Open Standards for Document Formats

The concept of a standard in ICT is well understood. According to Berkman (2005) they are the "mortar holding interoperable ICT systems together." Standards enable interoperability between diverse systems. Add to this the concept of openness, and

the concept is less well understood. Krechner (2005) suggests ten important rights that enable open standards, covering everything from IPR to how the standard is developed. However defined, open standards are increasingly recognised as central to interoperability – and have been credited with making the internet revolution possible.

Two of Krechner's 'ten rights', access to documentation and free usage of open standards, are often considered as the most important features. Perhaps the most important effect openness is that it encourages free competition and thereby diversity, which in turn protects against reliance on one product or platform. In other words, open standards lower risk. As put by Bird (1998): "(an) open standard is one which is used as a basis for producing interoperating products from a large number of providers – who can compete on any of a multitude of competitive advantages to the market buying their product."

The primary purpose of open standards for document formats is to make documents independent of the systems which generated them. This is of paramount importance for any organisation wishing to promote long term accessibility, including interoperability. Otherwise, in the worst case, a specific tool must be purchased and maintained in order to access an organisation's data; and this tool may well not be available on all ICT platforms. A separate advantage of open standards for document formats is that they act as enablers of fair competition in the marketplace, encouraging the development of tools which can compete because of their ability to interchange documents. The two most cited standards in government contexts are PDF/A and ODF.

PDF/A is an ISO standard for using PDF format (see www.pdfa.org). It is designed for the long-term archiving of electronic documents, and is currently based on PDF Reference Version 1.4. It restricts PDF in order to better guarantee prospects for archiving; in particular, it ensures visual reproduction of the document (PDF/A-1b compliance) and document structure, to allow searching and reuse (PDF/A-1a compliance).

Open Document Format (ODF, docs.oasis-open.org/office/v1.2/) is a standard developed by the Organization for the Advancement of Structured Information Standards (OASIS), and also published as an ISO/IEC standard. It is an XML-based format specification for office applications, including text and spreadsheets amongst others. It has gained traction as an open standard adopted within the open source OpenOffice.org application suite.

3 Research Approach

The research question addressed through this study is the following. To what extent are local government documents preserved electronically for discovery and re-use? In particular, we address three related sub-question. To what extent are official records available digitally? And, of those available, to what extent are they accessible, that is can be opened with the latest versions of current applications?

Importantly, of those which are available and accessible, to what extent are they searchable? In essence, a searchable document is available for re-use. This clearly rules out scanned PDF documents as reusable assets.

The question is made easier to answer in Sweden, which has a very strict policy on governmental responses to questions: all questions must be responded to.

We emailed a questionnaire to each municipality (290 in all) requesting minutes of selected executive board meetings. These minutes represent the decisions of the most senior board in each municipality, and hence need to be preserved. In fact, these minutes are archived and kept indefinitely (over hundreds of years) – usually in paper copy. Although rules on the long term electronic preservation of official documents are under consideration by the National Archives, there is currently no official policy.

Three requests were made. The first was for the minutes of the last meeting in 2008. The second was for the minutes of the first meeting in 1999; this was the oldest which could be expected in some cases as municipalities are allowed to selectively but systematically purge certain less important records after 10 years. Some municipalities interpret this as including electronic versions of minutes. The third was for the oldest minutes available electronically. The requests were sent in late January 2009 to the official email address of the registrar for each. It the request was not answered then a final reminder was sent in mid June. Responses to each email were recorded, together with all attached minutes.

The request was specifically for documents as currently stored electronically. It was made clear that the documents sent should not be specially created from physical archives or transformed from their stored format.

The study resulted in both quantitative and qualitative data. Quantitative data was analysed to answer the three specific questions on the extent to which minutes are available, accessible and searchable. The text of email responses was analysed qualitatively, to give some indication of factors affecting availability.

4 Quantitative Analysis

Of the 290 municipalities contacted, 267 (92%) have responded to date. Of these, 264 were able to provide documents. Of the three that did not, one does not save any electronic copies; another will start in 2009; the third only stores scanned documents and understood that these were not of primary interest to us. Of the respondents, only 88% responded promptly. This may be significant, as the request for 1999 was at the edge of a 10 year window and a significantly delayed response could affect availability due to deletion policies. The oldest available document gives some insight into the possible significance of this. In the worst case, 4 authorities may have removed the requested document during the period of data collection, sending a later 1999 document as the oldest. However, all of these responded early so this is a low probability.

About the documents received

Table 1 details the documents and formats sent by municipalities. In that table, "Other" includes files with the following extensions: .htm(6), .pro(4), .wpd(3), .tif(2), .dot(1), .027(1), .rft(1) and no extension (4). In all cases except .htm, .tif and .rft it was possible to open the document as an .rtf file. RFT is an (outdated) IBM binary format used on their mainframes, requiring special software to open it. In several cases, the filename extensions did not correspond to the actual format of the file which caused some confusion (e.g. some files were sent without extension, and some with a .doc extension were actually formatted as RTF).

Table 2 details the applications used by the municipalities for generating the documents. In analysing each file it was possible to identify which application had been used to generate it with a high degree of certainty. For almost all files sent to us in PDF we were able to identify the application used. Interestingly, no municipality that used OpenOffice.org sent us their minutes in ODF format.

We also looked at the application which generated each PDF file. Table 3 shows the results.

Table 1. Responses by document format

Format of response	Minutes from 2008		Minutes from 1999		Earliest Minutes available	
DOC	57%	(154)	74%	(120)	66%	(175)
PDF	38%	(104)	18%	(29)	24%	(64)
RTF	3%	(9)	6%	(9)	4%	(10)
DOCX[1]	1%	(2)	0%	(0)	0%	(1)
Other	1%	(2)	2%	(4)	5%	(14)
TOTAL[2]	100%	(271)	100%	(162)	100%	(264)
Paper only		(3)		(106)		(3)
Did not reply		(23)		(23)		(23)

Table 2. Application used to generate documents

Application generating the document	Minutes from 2008		Minutes from 1999	
Word	89%	(242)	96%	(153)
WordPerfect	0%	(0)	1%	(2)
IBM DisplayWriter	0%	(0)	0%	(0)
OpenOffice.org	1%	(3)	0%	(0)
Scanner	10%	(26)	3%	(5)
TOTAL	100%	(271)	100%	(160)

[1] One of the documents from 1998 was sent in .docx format, which was not available at the time of the meeting (see next section for related discussion).

[2] Up to 3% of municipalities provided their minutes for a given year in both DOC and PDF formats.

Table 3. PDF versions, 2008 and 1999.

Adobe PDF version number	Year	Minutes from 2008		Minutes from 1999	
1.2	1996	3%	(3)	14%	(4)
1.3	1999	20%	(21)	21%	(6)
1.4	2001	61%	(63)	38%	(11)
1.5	2003	10%	(10)	14%	(4)
1.6	2005	7%	(7)	14%	(4)
TOTAL		100%	(104)	100%	(29)

Clearly, if a document format was unavailable at the time that the minutes of a meeting were created, then reformatting has subsequently taken place. This is very evident with PDF files for 1999. Clearly, at least 66% of the PDF files sent for the year 1999 are not contemporary.

We further analysed documents for signs of post-facto curation. In particular, we excluded PDF documents created by scanning paper copies from the archive. We also noted documents created significantly after the date for the meeting, implying that the document was re-saved after further processing (for example, generated as PDF for a more recent initiative to place documents on the web).

Maintaining electronic access to documents is of significant and ongoing concern to archivists. Two primary methods are used when document formats are no longer supported: reformatting and emulation. In the former, documents are transformed to a current format and re-archived. In the latter, access to old formats is maintained by emulating the applications which originally produced them. Looking at the DOC formatted documents, there is clear evidence of reformatting – the most extreme case being the 10 year old minutes in .docx format.

Availability

To consider availability, we analysed the oldest document provided by each municipality. First we analysed each of these documents irrespective of its format and how it was created. When viewed cumulatively (Figure 1), it is possible to gain a sense of the level of electronic availability of documents by year over all municipalities. The data is not fully accurate. As our request only concerned one specific meeting (in 1999) it is evident from the comments in the responses that several municipalities also have other gaps, for various reasons, in what has been kept in electronic form (see further next section). We are in fact aware of some drop-outs – cases in which an older document than for 1999 has been provided but not one for 1999 itself.

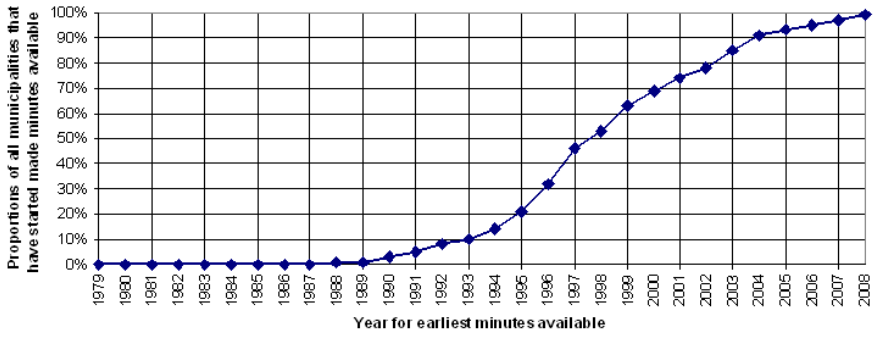

Fig. 1. Earliest available document by year

Accessibility

For accessibility we looked at both the 1999 and 2008 documents, but were particularly interested in 2008. This can give an indication of the extent to which e-government initiatives have penetrated local government, specifically in areas which are not directly controlled through legislation.

We classify a word processing document as accessible if it can be opened in the latest, currently available version of an appropriate word processing application in a manner which preserves its content and layout (to the extent that it can be read as intended, so for this analysis we ignored detailed formatting issues). We consider both proprietary and Open Source options. The only clear proprietary candidate was Microsoft Word (hereafter referred to as Word), as all word processing documents from 2008 were produced in Microsoft formats.

Table 4 shows the Word versions used in 2008 and for 1999 (for comparison). Of the total of 156 files received for 2008 two were docx-files and the other 154 were documents in "Word.Document.8" format. For the minutes received from 1999, 64 documents were supplied in "Word.Document.8" format and 56 were supplied in "Word.Document.6" format.

Table 4. Word versions used for 2008 (and 1999 for comparison)

MS Word version	Minutes from 2008		Minutes from 1999	
MS Office Word for Windows 95	0%	(0)	28%	(34)
MS Word 6.0	0%	(0)	24%	(29)
MS Word 8.0	3%	(4)	32%	(39)
MS Word 9.0	18%	(28)	0%	(0)
MS Word 10.0	17%	(27)	6%	(7)
MS Office Word	63%	(95)	9%	(11)
MS Office Word (for docx)	1%	(2)	0%	(0)
TOTAL	100%	(156)	100%	(120)

We classify any other document as accessible if it can be opened in the latest, currently available version of an appropriate application. Again we considered both proprietary and Open Source options. As all but two of these documents are in PDF, we conducted the tests with Adobe Reader 9.1 and Sumatra (Beta) v0.9.4. The two remaining documents were in .htm and .tif formats.

In testing the documents there were three primary concerns:

1. can the document be opened with the appropriate proprietary software application?
2. can the document be opened with the appropriate Open Source software application?
3. is the document searchable in each application?

The first step was to attempt to open all of the set of 'oldest' documents, in appropriate current applications. This is a test of their current accessibility. Clearly reformatting may have improved or compromised accessibility, and may have compromised reusability (for example if a scanned paper copy was maintained).

We initially considered the level of success in opening word processor documents using Word 2007 and OpenOffice.org 3.1. We considered the oldest documents first. It is documented on the vendor website that early versions of the .doc format are not supported in recent versions of their software (Word 2003 and Word 2007). This was confirmed when we tried to open documents stored in Word for Windows 1.x (1 document) and 2.0 formats (6 documents in all). Interestingly, all documents, including these, could be opened in OpenOffice.org 3.1. However, the Windows 1.x document was treated as a text file. The text of the minutes could be discerned, but most of the formatting was lost and the binary encodings were presented as unicode text. Both applications were able to open all other documents with file formats .doc or .rtf. Both were also able to open a number of documents with other file formats, namely WordPerfect 5.x and 6.x and the single occurrence of a docx file (clearly curated).

All documents from the 2008 set could be opened in both Microsoft Word 2007 and OpenOffice.org 3.1, although a number of problems were still encountered. These included problems related to: formatting that breaks; macros and the use of different variables and templates that are not kept embedded in the document; locked and password protected documents; documents that try to access embedded SQL-queries (with dependencies on other applications); and Word comments that cause problems (the documents opens with comments that are generated in a previous version of Word).

All PDF documents could be opened with Adobe Reader 9.1 and with Sumatra (Beta) v0.9.4. Although we were able to open all files there was for one PDF-file a font problem identified when using Adobe (resulting in an error message which warned that the file may not display and print correctly), whereas such an error was not indicated when we opened the same file using Sumatra. We therefore concentrated on the question of whether the document was searchable – this being an indicator of whether a document was reusable.

Reusability

For studying reusability, we are interested primarily in the information within a document, not its styling. We consider the data sets from early 1999 and late 2008, giving us a view of two time points roughly 10 years apart.

Looking at the minutes submitted of the first meeting of each municipality held in 1999 (29 documents in PDF format), we found that 5 (17%) were produced by some form of scanning and were not searchable. All of the other documents were searchable.

In looking at the details of each non-scanned document, it could be seen that the PDF had been generated using a variety of versions of Word. Considering these together with the .doc documents from the same period, we obtain a snapshot of the latest versions of Word used to save each document (see Table 5). This does not necessarily reflect the Word version used to create the file, since PDF generation may have been done at a later point in time, and the document may have been opened and saved in a later version of Word at a later date – for example to update its format.

Finally, we considered the documents supplied from 2008. The majority of these were .doc format (see Table 1), but the proportion in PDF format had significantly increased (from 18% to 38%). This is unsurprising, as many municipalities now routinely publish meeting minutes on the web, but few extend back as far as 1999. Some municipalities even supplied URLs in their response rather than attaching documents.

The increase in use of PDF for maintaining electronic copy of documents raises further questions. In particular, are the documents in a form of PDF which is reusable, accessible and open? For this research we checked whether each PDF file could be searched (a necessary prerequisite for reusability) and whether it was open (in a format compatible with PDF/A, and in particular the weaker requirement PDF/A-1b). PDF/A is a subset of PDF designed to be more suitable for the long-term archiving of documents. Disappointingly, only one document from 2008 was found to be compliant with PDF/A-1b.

Of the 104 PDF-documents from 2008, the percentage non-searchable was found to be 23%, whereas for the 29 documents from 1999 the percentage non-searchable was 17%. It should also be noted that not all non-scanned documents were searchable. Interestingly, one of the documents was generated by OCR using the Adobe Acrobat paper capture plug-in.

5 Qualitative Analysis

Analysing the document attachments gives insight into the extent of availability and accessibility of files. However, any explanation of why certain documents are unavailable or inaccessible requires a qualitative study. For this, we analysed the content of the emails to which documents were attached. In most cases an

explanation was given if there was any problem in meeting the request. We present these for insights into availability (and hence archiving policies), and also, where relevant, for accessibility.

In 40% of cases municipalities could not provide the requested minutes from 1999. Reasons varied. At one extreme, there was a practice of keeping no electronic archives. In some cases, there was a policy of periodically pruning back. For example:

"Electronically stored copies are selectively deleted and this has been the case with the minutes from 1999"

Such policies are not always strictly systematic. The minutes of these meetings – although related to the top level committee of the municipalities – are not protected by law except in their paper form. Hence this next response:

"The minutes written before 2000 have probably been deleted at the time the minutes were signed and archived. There is no record kept about how this was handled exactly in the year 1999."

This lack of obligation to protect electronic copy of minutes has led to a lack of policy, allowing this informal approach. This potentially impacts both on availability and accessibility. Hence, even where minutes are made available this seems to be as a result of other informal processes. For example,

"(The municipality) does not have any electronic storage of documents ... . However, the minutes are temporarily available as PDF and stored in an ordinary folder on disc (and) used for presentation on our home page."

Most had a less systematic reason for unavailability. Most pertinent to our enquiry is loss unavailability due to legacy systems and formats:

"I can unfortunately not find anything from 1999. A different system was in use then, which we do not have access to today."

"Unfortunately we do not have the minutes from 1999 for technical reasons."

It is clear that even after only ten years there are problems related to accessing or interpreting files which are known to exist. In a number of instances this was because the file uses a proprietary format which is only interpretable by the legacy tool which created it. This may imply extra cost and delay in meeting a request:

"The oldest minutes are not available and the minutes from 1999 should be available but the tool ... has been phased out from the organisation. Your request has triggered our IT department to resurrect the software."

In fact, it may be significant enough for the organisation to seek a way of not incurring the expense:

"From 1994-08-10 we used Ergo-ord for electronic documents. We have not been able to recreate the first document from 1993-01-13, a specific environment is necessary for this which we have not set up. You will have to be satisfied with a document from 1994-08-10, which is the first Word version. Please come back to me if you insist on the minutes from 1993."

In other cases, it was not possible to completely reconstruct the document even though the organisation was willing to make the effort. In one case this resulted in significant data loss:

"For the documents from 1990 I have been forced to ask the IT section to convert so that the documents can be read. The meeting minutes from 1990-01-09 were inserted into a pre-printed template, hence this is missing in the document. The minutes from 1999-01-12 contain only 2.5 paragraphs and 4 pages. According to the original minutes in paper form there should be 35 pages."

In a different case, layout information was lost even for recent minutes, and the older minutes also suffered data loss (in the response a paragraph refers to a distinct item on the agenda):

"The minutes from 2008 are digitally preserved in Word but in that the paragraphs are not in the correct order. The 1999 minutes are only partially preserved: the first page and a few of the paragraphs are missing" ... "We had the principle of one Word file for each paragraph at the time, so it is messy to organise these.

Such access problems were not limited to documents generated by tools which are no longer used. In a number of instances old files, in a proprietary format, could not be opened natively in the latest version of the tool which created them. Interestingly these same files could be opened natively in an open source tool, OpenOffice.org.

6 Discussion and Conclusions

There is no evidence from our study that municipalities have a data curation policy with respect to executive minutes. In the absence of a direct duty to preserve electronic copy, curation is left to the work practices of individuals.

Where electronic copy is kept, proprietary and closed formats are overwhelmingly used for public documents, even though there is experience of losing access to, or increased cost of access to documents because of formats which are no longer supported. Further, and perhaps more significantly, we find no evidence that this situation is changing.

Our general finding is that availability of electronic copies of executive board minutes for municipalities in Sweden is very variable, and accessibility is poor. In particular, there is little evidence of the existence of policies to maintain electronic copies of documents, and little awareness shown of open standards and their importance in data curation. It is striking that no municipality provided a document in a reusable, open standard document format. This stands in stark contrast with stated central Government policies. The study also highlights a consequent lack of strategies in organisations regarding effective communication and archiving of electronic data.

As a result, there are already many gaps in the electronic data record even for the most recent 10 year period. In this very real sense, the mooted digital dark hole in public records is fast becoming a reality.

References

BBC (2007) Warning of data ticking time bomb, BBC News, 3 July, http://news.bbc.co.uk/2/hi/technology/6265976.stm, accessed 2009-12-23

Berkman (2005) Roadmap for Open ICT Ecosystems, Berkman Centre for Internet & Society at Harvard Law School.

Bird, G. B. (1998) The Business Benefit of Standards, StandardView, Vol. 6, No. 2, June/1998, pp. 76-80.

DCC (2005) Digital Curation Manual: Instalment on 'Open Source for Digital Curation', 1 August, http://www.dcc.ac.uk/resource/curation-manual/chapters/open-source/

EU (2004) European Interoperability Framework for pan-European eGovernment Services, European Commission, Version 1.0, http://ec.europa.eu/idabc/servlets/Doc?id=19529

Krechmer, K. (2005) The Meaning of Open Standards, In Proceedings of the 38th Hawaii International Conference on System Sciences – 2005, IEEE Computer Society, Los Alamitos.

Regeringen (2009) Kommittédirektiv: Delegation för e-förvaltning, Dir. 2009:19, 26 March, http://www.sweden.gov.se/content/1/c6/12/40/02/ec50b88b.pdf *(in Swedish)*.

Regjeringen (2009) Nye obligatoriske IT-standarder for staten vedtatt, Fornyings- og Administrasjonsdepartementet, Pressrelease: 2 July, http://www.regjeringen.no/nb/dep/fad/pressesenter/pressemeldinger/2009/nye-obligatoriske-it-standarder-for-stat.html?id=570650 *(in Norwegian)*.

SOU (2009) Strategi för myndigheternas arbete med e-förvaltning, Statens Offentliga Utredningar: SOU 2009:86, e-Delegationen, Finansdepartementet, Regeringskansliet, Stockholm, 19 October, http://www.sweden.gov.se/content/1/c6/13/38/13/1dc00905.pdf *(in Swedish)*.