Kragen Javier Sitaker
Rivadavia 2318
Piso 4 Dept. B
Capital Federal
C1034ACP
Argentina

February 18, 2009

Mr. Stephen Gura
Deputy Associate General Counsel
Mr. Mark Shonkwiler
Assistant General Counsel
Federal Election Commission
999 E Street, N.W.
Washington, DC 20463

Via email: agencypro2008@fec.gov

Dear sirs:

Attached please find my comments pursuant to the Federal Election Commissions Notice of Public Hearing and Request for Public Comment, posted in the Federal Register, Vol. 73, No. 236 on December 8, 2008.

As a US citizen living abroad, my primary concern is with the interests of the public, not with the interests of the regulated community. While I recognize that transparency, effectiveness, and fairness in enforcement proceedings are a *sine qua non* of the FEC's mission, it is my view that they are not sufficient; rather, the Commission also has an obligation to provide the information gathered through its enforcement activities to the public in order to serve the purpose for which the Commission was created: ensuring the fairness of our Nation's electoral process, and allowing the public to take political action to correct problems in the process.

My work with the FEC's data has been as a consultant under contract with Watchdog.net, LLC, which is funded by the Sunlight Foundation. However, these comments are on my own behalf; I am not representing either of these two organizations.

Thank you for soliciting our comments on these important issues.

Sincerely,

Kragen Javier Sitaker

Comments to the Federal Election Commission

We thank the Commission for requesting public comment on its policies and procedures, particularly with such a broad ambit.

Our expertise is limited primarily to the Commission's disclosure of filing information, and our comments are accordingly limited in scope.

Table of Contents

Importance of Disclosure; Context

The FEC was established in 1975 to stop certain financial abuses of the federal electoral system, and to justify and reestablish the public's confidence in that system. Accordingly, public disclosure of funds raised and spent to influence federal elections is a key part of its mission, and the Commission already publishes a great deal of data on its World Wide Web server for that purpose.

However, this disclosure falls short of the ideal, in a number of ways which substantially limit public accessibility of the data. The Sunlight Foundation's comments describe some of the shortfalls, and they are substantially correct.

Since the Commissioners and their staff spend most of their time taking enforcement actions and interacting with members of the regulated community, it is not surprising that the Commission's request for public comments concerns primarily issues of enforcement and issues of interest to the regulated community rather than issues of interest to the public at large. However, we would like to request that the Commission also start to take its public disclosure mission seriously. The Commissioners' comments on this matter in the January hearing were very encouraging.

The Sunlight Foundation's comments discuss the importance of a user-friendly web site for searching and viewing FEC disclosures. Mr. Clay Johnson's testimony in the January 14th hearing says, “Your number 1 priority in fulfilling your mandate to publicly disclose campaign finance information should be to provide high-quality and accurate data to citizens in a way that is comprehensive and understandable.” He is correct.

These are both important goals, but as Mr. Johnson emphasizes in his testimony, “understandable” is of secondary importance to “comprehensive,” that is, complete and documented. As Mr. Johnson explained, there are a number of web sites that already provide interfaces for querying and reporting on the data, using the downloadable data files. It is much easier for a member of the public to build a user-friendly web site, if they can download comprehensive data, than for a member of the public to extract comprehensive data from a web site built only to be user-friendly.

Accordingly, our comments focus primarily on what is necessary for the data available to the public to become comprehensive.

Procedural Aspects of Filing Disclosure

There are a variety of nontechnical measures that would substantially improve public accessibility of filing data.

A. Old Forms, Schedules, and Instructions

The original electronic filings from previous years constitute an important public record, both because of the problems with the COBOL master data files described in the Sunlight Foundation's written comments and testimony, and because of the general principle that public records should be available to the public in the form in which they were originally created.[1] Unfortunately, the filings are not self-contained; the data in them represent answers to specific questions on FEC forms and schedules, and must be interpreted in light of the forms and schedules and their corresponding instructions.

Proposal

These forms are available from the Commission's web site, but apparently only in their current forms. For reasons of public access, we propose that the Commission maintain a web-accessible archive of *all* old versions of each form, schedule, or set of filing instructions, with information about when they were valid. This should represent a minimal administrative burden, since presumably fewer than 40 versions of each have been created, and none of them should be difficult to find a copy of.

Additionally, for the benefit of members of the public who are reading old versions of these forms, the Commission should maintain on its web site a chronological list of changes to all of these artifacts — if not since the inception of the Commission, at least for recent years. Presumably this list of changes already exists in some form, however fragmentary, because it is needed by filers as well as those seeking to understand old filings. Without such a list of changes, a reporter or other member of the public needs to wade through an overwhelming amount of repetitive information in order to understand a series of reports from several previous years.

B. Old File Formats

The Commission documents the electronic filing format on its web site[2] — however, this information is designed for vendors of electronic filing software, not members of the public seeking to exercise their right to access public information. Accordingly, it only includes full documentation for the latest format. Over the years, the filing format has changed a number of times (the current filing format is Version

[1] Of course, this does not imply that the data should not also be available in more convenient or consistent formats.
[2] http://www.fec.gov/elecfil/vendors.shtml

6.2.1.2 of the format), but apparently the old filing formats are not documented on the Commission's web site, nor is there even a list of old filing formats.

There is a limited amount of information in the vendor information file about recent obsolete format versions, going back only to version 5.3. Through other channels, we have been able to obtain documentation of old filing formats going back to version 3, but we have no way of validating that the documentation is authentic, and we have no documentation at all for filing format version 2.

Without this information, it is very difficult to make sense of old electronic filings, and information obtained from guesses at their formats must be treated as unreliable.

Proposal

We request that the Commission consider the right of the public to access historical filing data to be as important as the need for vendors to produce filings in the current format, and accordingly that the Commission provide a list of all historical file formats on its web site, linked to complete documentation of those formats. As with old versions of the forms and instructions, this information already exists; it is just a matter of finding it and putting it online.

C. Filings Not Originally Filed Electronically

Original filings from before the year 2000, and some filings since then, are not online, because they were not originally filed electronically. For these filings, we have only the processed COBOL data, which is known to differ from the information in the original filings in any number of ways, such as field truncation and omission of pre-amendment filing information.

Proposal

We request that the Commission make these filings available online if they still exist in some original form, which may involve scanning a large amount of paper, and that the Commission minimize the number of future filings that are not filed electronically. Although this is a larger project than finding the few dozen pages of old file format documentation and the few hundred pages of old forms and instructions and putting them online, it should still be feasible for around a penny per scanned page.

D. Correspondence from the FEC to Regulated Entities

Although the Commission has *ex parte* rules that prohibit covert contacts between Commissioners or senior staff members and entities under investigation, there are a large number of contacts between the Commission and regulated entities, even when those entities are not under any kind of investigation. Many F99 filings with the Commission consist of textual responses to such contacts. In principle, these correspondences between the Commission and regulated entities are a matter of public record. However, only the filings by the regulated entities appear to be available on the Commission's web site. The consequence is, in effect, that the public can only hear one side of the conversation.

Proposal

We propose that the Commission make all such correspondence available, with the date and FEC ID of the correspondent in machine-readable form.

Technical Aspects of Filing Disclosure

A great deal of effort has clearly gone into the Commission's electronic filing software, but despite this, technical aspects of filing disclosure pose many unnecessary problems.

A. Data Formats

According to the January testimony, the Commission is currently planning to adopt new data formats based on modern standards such as XML. This is a good idea, but it will not necessarily solve the data format problems even for new data, and of course it cannot change the format in which electronic filings from years ago were submitted.

The current data formats are unnecessarily complex and error-prone. Mr. Johnson and Ms. Miller's comments have already discussed the problems with the COBOL data format — for example, that it is fixed-width, and uses a surprising encoding for negative numbers — so we will discuss some examples of problems with the filing formats.

All of the filing formats are positional rather than name-value oriented; the 33rd field on a given line has a certain meaning which is not necessarily related to the 32nd or 34th fields, and which must be looked up in a separate document. This process is error-prone and unnecessarily difficult to reverse-engineer. These formats take up less space than name-value formats, but even for high-volume filers such as MoveOn and Obama, the size is not significant with modern technologies.

The filing data formats before version 6.1 were comma-separated, with the exception of a few lines that were not comma-separated, such as those representing file headers in versions before version 3.0, and those representing text attachments. This creates the complication that standard libraries for comma-separated data files have difficulty handling them. XML or JSON would avoid this problem.

In versions prior to 5.1, the format included fields for people's names that were internally structured by separating parts of the name with a "`^`" character. However, the format included an option to separate parts of a person's name with another character, which increases the complexity of software that deals with data written in the format for no increase in power. For example, filing 31454 uses the "`>`" symbol to separate parts of names; however, filing 33818 claims to use the character "`0`" to separate parts of names. In fact, the names in this filing are separated with "`^`"; for example, "`Allyn^Margaret^Ms.`". The FEC appears to have accepted this filing.

In other cases (for example, filing 23422) instead of the usual maximum of four "^"-separated segments to a name, we find as many as 32, most of them empty. The FEC appears to have accepted filing 23422 as well.

Version 6.1 uses a control character, instead of a comma, to separate data fields, avoiding the need for quote marks. This makes the file harder to process in some ways (for example, some programs for handling text decide that the file is not text and refuse to display its contents; and the particular character chosen is erroneously coded as a line separator in the Unicode standard) but does simplify the format.

Finally, in all of the format versions before version 6.1, according to the specification, an amount field coded as “`100`” means $1.00, not $100.00. The latter amount is to be coded as “`100.00`” or “`10000`”. Given this bizarre design choice, it is not surprising that there are a large number of filings in which contributors appear to have donated $1.00 or $2.00 to a PAC or candidate committee.

A simple conversion of the format to XML (or JSON, or SQL, or BER, or RFC-822, or S-expressions, any of which would be preferable to the current ad-hoc CSV dialect) would not solve most of these problems, except for the problem of commas. A poorly-designed XML format could define that `<amount>100</amount>` encodes a dollar amount of $1.00, or that a candidate name of “Joseph Biden” should be coded as `<name>Biden^Joseph</name>`.

B. Data Format Documentation

Although the Commission has highly-skilled technical experts on staff, it appears that the Commission did not consider the task of designing and documenting the electronic filing formats sufficiently important to warrant the attention of these experts. This is a mistake. Without clear definitions of the syntax and semantics of these filing formats, filers and filing-software authors will make errors that change the meaning of their filings; and members of the public afterwards will not be able to ascertain the intended meanings of the filings with any certainty.

The avoidable mistakes in the design of the filing formats described in the previous section would not be nearly as serious if the documentation of the filing formats were clear, readable, and unambiguous, and corresponded to how the filings were actually processed. However, the format documents contain errors and ambiguities at every level, from simple spelling errors (“carridge-return”, “hexidecimal”, from FEC_v520.doc) to major conceptual omissions.

By way of example, here is a list of some of the biggest unanswered questions, mostly from FEC_v520.doc.

- If “fields may not begin with blanks”, as it claims on p.2, why are there so many electronic filings containing fields that begin with blanks?
- What is the algorithm to construct an amended filing, given an original filing and an amendment? Do amended schedules simply replace the original schedule with the same `tran_id`, or are they merged somehow?
- In one of the 6.x format documents, it is explained that `tran_id` is becoming case-insensitive — that is, that `tran_id`s that differ only by letters in one being uppercase where they are lowercase in the other, will be considered equivalent in the future. When did this change become effective? Could it change the meaning of amendments?
- P.9 says there is an example of a header record. Where is the example?
- P.9 references the “Rpt Id” of the original report. What is the format of this Rpt Id, and where is it obtained? Different filing software seems to format this field very differently.
- Only printable ASCII characters will be accepted, according to p.3. What about filing 181941, then? It contains text encoded in Windows-1252. There are other filings that contain non-ASCII text in other encodings or character sets. How are they to be interpreted? Why were they accepted by the Commission?
- What exactly is the delimiter that ends a free-form text section in an F99 filing? The document claims that it ends at an “[ENDTEXT] record” or an “[EndText] record”, and provides an example, but never defines the delimiting record clearly. In accepted filings, we have seen “`[ENDTEXT]`” on a line by itself, “`[END TEXT]`” with a space, “`[ENDTEXT]`” with a quote character after it, “`[ENDTEXT]`” on the end of a line after hundreds of characters of text, and several other variations.

In format version 6.2, there ASCII-only requirement is loosened; there is some text written under the misapprehension that ASCII (a code that was standardized in 1963 and updated in 1967, and is used to represent nearly all English text in nearly all computers) has code points past 128. Unfortunately, the description of the allowed characters and code points does not match any character set we know of: not ISO-8859-1, not

CP857, not Windows-1252, not Unicode, not UTF-8. The reader is left to guess which encoding is meant, and different readers will presumably guess differently.

There are well-known rigorous formalisms for file formats dating back to the 1950s, such as Backus-Naur form and context-free grammars. They are well-understood, and everyone who has read a standard for a programming language such as C, Fortran, or COBOL has seen one. There are software tools to automatically analyze the structure of a file, given a BNF or CF grammar for the file format. Such a grammar would be a small fraction of the size of the FEC file format specification document and would leave no doubt about the intended structure of the file. They do not explain semantic issues like some of the problems cited above, but the syntactic ambiguities cited above are entirely unnecessary.

Proposal

We propose:

- that the Commission document how each of these ambiguities is resolved inside of its own systems, and disclose any communications with software vendors clarifying these issues;
- that the Commission define a new filing data format without the unnecessary complexity of the original filing formats, and which is flexible enough to be used into the future without backwards-incompatible changes;
- that it document this format properly, with as little ambiguity as is practical;
- that it require new electronic filings to be in this format;
- and that it write and publish software to translate all old filings to this new format.

However, transparency dictates that the old versions remain available as well, so that errors in the translation process can be uncovered.

C. Software

The Commission provides software known as FECFile and FECheck to electronic filers. FECFile is an interactive program for creating electronic filings, while FECheck verifies that an electronic filing is properly formatted, whether that filing was created by FECFile or by some other software. The Commission also appears to use FECheck to verify that electronic filings are in the correct format and automatically reject them if not.

These programs are written by employees of NIC Technologies under contract with the Commission, and are apparently not used by any of NIC's other customers; in particular, NIC does not appear to use its copyright in this code to provide value-added FEC filing software to filers. The public funded the development of these programs, and the integrity of our elections depends, in part, on the correct processing of information through them. Despite this, the only version of the programs that seems to be publicly available is a “compiled” format in machine code for a single kind of computer running a single operating system.

This limits the transparency of the process, requires vendors of electronic filing software to duplicate development work already performed and paid for by their tax dollars, and requires electronic filers to obtain and maintain a computer running that operating system, paying that operating system vendor for the privilege of participating in our political system.

The duplication of effort is not merely a waste of money for the vendors. It also results in each vendor interpreting the format specifications in different ways, introducing unnecessary variation in the format of electronic filings. This makes it more likely that the Commission's software, or other software written to interpret publicly disclosed electronic filings, will interpret an electronic filing differently than the filer intended.

From a casual glance, FECheck at least appears to be written in Micro Focus COBOL. Although COBOL has its technical disadvantages, it has the advantage of being well-standardized, widely-known, and highly human-readable, so the source code would be very useful.

Proposal

We suggest that the Commission negotiate with NIC to obtain full human-readable source code to the current and all past versions of these programs under an “open source” license approved by the Open

Source Initiative, and make that code publicly available. This would enable the public to inspect how they work (and how their interpretation relates to the written English format specifications), and it would enable electronic filing software vendors to reuse the work already done where it is applicable.

Sometimes negotiations like these are complicated by a secondary revenue stream from non-governmental customers, who buy licenses from the original software vendor for an "enhanced" version of the software, but NIC does not seem to have any such secondary revenue stream.

Web Site Searchability and User-Friendliness

As we have said, improvements to the structure of the FEC's web site are of secondary importance to improvements to the available data.

Proposal

However, if the Commission chooses to devote resources to improving the user-friendliness of its web site, we concur with the recommendations in Ms. Miller's written comments and Mr. Johnson's testimony:

- provide a single search box for all data and group search results by type;
- provide RSS feeds;
- provide "box score" summary information;
- explain the technical language;
- provide JSON APIs designed in accordance with REST for third-party applications, or XML if JSON is not acceptable;
- use HTTP GET for searches.