# Decoding dat files from a Thermo Element™ ICP Mass Spectrometer

**John Hartman**
(Department of Computer Science, University of Arizona, Tucson AZ 85721 USA)

**Rob Franks**
(Department of Earth and Planetary Sciences, U.C. Santa Cruz, Santa Cruz CA 95064 USA)

**George Gehrels**
(Department of Geosciences, University of Arizona, Tucson AZ 85721 USA)

**Jeremy Hourigan**
(Department of Earth and Planetary Sciences, U.C. Santa Cruz, Santa Cruz CA 95064 USA)

**Philip Wenig**
(Lablicate GmbH / OpenChrom, Martin-Luther-King Platz 6 Hamburg 20146 Germany)

December 12, 2017

**Abstract**

We present the internal structure of the binary data files generated by the data acquisition software of a Thermo Element$^{TM}$ ICP mass spectrometer. This gives researchers access to the raw intensities measured during an analysis, rather than relying on the results produced by the Thermo "Result Display" software. In addition to describing the file format, we also provide a Python module called ExtractDat that when run as an application produces a CSV file of the information in a binary data file, and can also be used as a library to extract the information for use in other third-party software.

**Keywords:** Isotope Geochemistry, Decoding, Reverse Engineering, File Format, Thermo Element$^{TM}$

**1.1 Introduction**

The Thermo Element$^{TM}$ is a double-focusing magnetic sector ICP Mass Spectrometer that is capable of analyzing ppm- to ppt-level concentrations across a broad mass range. It is one of the most commonly used instruments for determining elemental concentrations and isotope ratios of geological materials.

The Element2 is able to measure signals with a high (10e9) dynamic range in large part because the detector system operates in two modes: a pulse-counting mode for signals that are less than 5M cps (counts per second), and an analog mode for larger signals. Essential for comparing signals measured in the two modes is the Analog Correction Factor (ACF), which is used to convert an analog signal into pulse intensity. The ACF is determined by measuring the intensity of a signal in both modes between 50,000 and 5,000,000 cps and calculating a linear conversion factor. The accuracy of the ACF is critical for reliable determination of concentrations and isotope ratios given the common need to compare intensities measured in both pulse and analog mode. The Element XR is also able to measure signal intensities with a Faraday collector, which is calibrated to pulse intensities with a Faraday Conversion Factor (FCF).

One of the challenges of interpreting data from an Element2 is that the raw counts measured during an analysis are not directly accessible from the Thermo data acquisition software. Rather, the raw counts are saved in a binary "dat" file, which is then converted with the Thermo "Result Display" software into an ASCII-format "FIN2" file. Unfortunately, FIN2 files do not include individual integrations (only averages), and do not report analog-mode or faraday-mode intensities that are below 5M cps. This critical information is not available from any of the Thermo software systems, and precludes detailed examination of the measured intensities, ACF's, and FCF's.

In an effort to address this challenge, we reverse-engineered the binary dat file created by the Thermo data acquisition software and developed a decoding routine that extracts raw intensities from dat files. What follows is a description of the dat file format and the method used by our dat-file decoding software.

**2.1 Dat file format**

The raw counts from the Element are stored in a binary dat file. The individual data items vary in size. Within this document we shall use the term "nibble" to refer to a 4-bit quantity, "byte" to refer to an 8-bit quantity, "short" to refer to a 16-bit quantity, and "word" to refer to a 32-bit quantity.

The file has the following format.

```
File Header
Scan index
Scan 0
        Scan Header
        Mass 0 Intensities
                Intensity 0
                        Pulse
                        Analog
                        Faraday
                Intensity 1
                        Pulse
                        Analog
                        Faraday
                …
        Mass 1 Intensities
                Intensity 0
                …
        …
Scan 1
…
```

**2.2 File Header**

Every dat file starts with a header block consisting of 89 words that contain information about the contents and format of the rest of the file. The meaning of many of these fields is unknown, however only a few are necessary for parsing the rest of the file. The known header fields are shown in the following table. The "Name" column is a name we've given the field for reference in this document. Units are given in parentheses and any notes appear in square brackets.

| Offset (bytes) | Size (bytes) | Name | Description |
|---|---|---|---|
| 148 | 4 | ScanIndexOffset | Offset of the scan index (bytes) |
| 172 | 4 | ScanIndexSize | Size of the scan index (words) |
| 176 | 4 | Timestamp | Experiment start time (seconds) [since epoch] |

Note that several things do not appear in the file header that one might naturally expect, such as the number of masses, identification of the masses, and the number of intensities per mass

in a scan. The number of masses and the number of intensities per mass are determined when parsing the rest of the file; the identity of the masses cannot be determined from a dat file.

## 2.3 Scan Index

The scan index appears in the file at the offset specified by the *ScanIndexOffset* in the file header. The scan index contains the offsets of the data for each of the scans.  The size of the scan index, and therefore the number of scans, is indicated by *ScanIndexSize*.

| Offset (bytes) | Size (bytes) | Description |
|---|---|---|
| 0 | 4 | Unused |
| 4 | 4 | Offset of Scan 1 data (bytes) |
| 8 | 4 | Offset of Scan 2 data (bytes) |
| … | 4 | … |

## 2.4 Scan Header

The intensities for a particular scan appear at the offset specified in the scan index for that scan and begin with a scan header. The scan header consists of 47 words, many of which are unknown. The known ones are shown in the following table.

| Offset (bytes) | Size (bytes) | Description |
|---|---|---|
| 28 | 4 | Time delta from previous scan (ms) |
| 36 | 4 | Scan number [starts with 1] |
| 48 | 4 | ACF  * 64 |
| 72 | 4 | Time delta of previous scan from start of experiment (ms) |
| 76 | 4 | Time delta of current scan from start of experiment (ms) |
| 124 | 4 | EDAC / 1000 |
| 140 | 4 | FCF *256 |

## 2.5 Mass Intensities

Following the scan header are the intensity data for the scan. The data are organized as a sequence of word-size records (32-bit quantities). The high-order nibble (bits 31-28) contains a *tag* field that identifies the *value* field stored in the remaining 28 bits. The tag and their meanings are shown in the following table:

| Tag | Description | Value field contents |
|---|---|---|
| 1 | Intensity data | See below |

| 2 | Magnet mass | Mass = value / $2^{18}$ |
| 3 | Channel time | Channel time (ms) |
| 4 | Accelerating voltage | Voltage = EDAC / 1000 / value / $2^{18}$ |
| 8 | End of mass | Duration (ms) |
| 11 | ?? | NA |
| 12 | B-scan | NA |
| 15 | End of scan | NA |

## 2.6 Intensity Data Record

Intensities are stored in the dat file using a custom floating point representation. Each intensity is stored in 4 bytes, the first two bytes of which are the significand and the next two bytes are an encoded version of the exponent. Both of these values are in little-endian byte-order.

The format of the record is shown in the following table:

| Bits | 31-28 | 27-24 | 23-20 | 19-16 | 15-0 |
|---|---|---|---|---|---|
| Contents | Tag (1) | Flag | Type | Exponent | Data |

The meanings of the different fields are shown in the following table:

| Field | Meaning |
|---|---|
| Flag | 0: measurement is valid |
| Type | 0: analog<br>1: pulse<br>8: faraday |
| Exponent | data should be scaled by $2^{exponent}$ |
| Data | Measured value |

The formulas for deriving the value from the record are as follows. Note that the ACF and FCF values are obtained from the scan header.

| Type | Formula |
|---|---|
| Analog | ACF / 64 * Data * $2^{Exponent}$ |
| Pulse | Data * $2^{Exponent}$ |
| Faraday | FCF / 256 * Data * $2^{Exponent}$ |

## 3.1 Issues with FIN2 files

Comparisons of decoded dat files and FIN2 files suggest that there is excellent correspondence for most values. Figure 1 shows the linear relationship of values extracted from FIN2 files and decoded dat files for ~10,950 measurements. The isotopes measured include Si29, P31, Sc45, Ti49, Y89, Nb93, La139, Ce140, Pr141, Nd142, Sm152, Eu153, Gd157, Tb159, Dy164, Ho165, Er166, Tm169, Yb174, Lu175, Hf177, Ta181, Hg202, Pb204, Pb206, Pb207, Pb208, Th232, U235, and U238. One integration was conducted on each peak for Si-Ta, four integrations per peak were measured on Hg, Pb, Th, and U isotopes. The analyses were conducted by LA-ICPMS on zircon crystals using methods described by Chapman et al. (2016).

In detail, as shown on Figure 2, there are discrepancies between FIN2-file and dat-file values for both low-intensity and high-intensity readings. Low-intensity offsets occur only for elements with multiple integrations, which include Hg , Pb, Th, and U. As shown on the Low Intensity Errors sheet in Appendix 1, the errors in FIN2 readings arise from omission of either one or two of the four measured values when averages are calculated. Errors occur in ~35% of the low-intensity measurements, with patterns as follows:

- The most common issue arises from omission of a zero value when it is in the fourth position, resulting in a FIN2 value that is less than the dat-file value. This error occurs in 11.6% of the low-intensity measurements.
- Omission of a zero value in the first position is also quite common (11.2% of measurements).
- Less common is the omission of a low-intensity value from the fourth position. In all cases the omitted value is less than the value in the first position, but can be larger or smaller than values in the 2nd and 3rd positions. This error occurs in 4.6% of the measurements.
- In ~4.2% of the measurements a low-intensity value in the first position is eliminated. In all of these cases the omitted value is less than the fourth value, but can be smaller or larger than values in the 2nd and 3rd positions.
- In 4.2% of the measurements, zero values in the third and fourth positions are omitted, and the FIN2 average is calculated from only the first two values.

Low-intensity values are apparently omitted due to an issue in the Thermo acquisition software, which arises from having a search window >0% with one or more zero count channels in the acquisition window. In some cases the search moves the integration window out of the acquisition window, so one (or more) of more channels is not included in the average. Changing the search window to 0% produces correct results.

In our application, the impact of omitting one or two of the measured values depends on how the intensities are used. If the measurements are on backgrounds that are subtracted from much larger peak intensities, e.g., 206-238 in our measurements, omission of one or two low intensity values has negligible impact on the background-corrected intensities. However, if peak

intensities are of low intensity, e.g., 202 and 204, subtraction of incorrect background values will lead to significantly underestimated peak intensities.

There also are discrepancies in high-intensity values (Figure 2; Appendix 1), in all cases with FIN2 files reporting lower intensities than dat files. These discrepancies have discrete values that correlate with omission of between 1 and 32 counts in the unscaled analog values that are used to calculate FIN2 intensities, although typically the number of counts omitted is a power of 2. Fortunately, all of these errors are negligible in comparison with the large magnitude of the measured signals.

**3.2 Improved ACF determination**

As noted above, having access to pulse-mode and analog-mode intensities enables calculation of an ACF value for each mass during an analytical session. Figure 3 (Appendix 2) shows ACF values calculated from measured pulse-mode and analog-mode intensities during a typical 3.5-hour analytical session using the acquisition parameters described above. Also shown is the average ACF value (available from FIN2 files) for this session. Importantly, average ACF values for different elements vary by ~7.4%, and are as much as ~5.6% lower and ~1.6% higher than the average ACF value calculated from all elements (and reported in FIN2 files).

These analyses also show the degree to which ACF values change during a single analytical session. As shown in Figure 4 (Appendix 3), the ~1.6% decrease in ACF during the session reported above (Figure 3) is not unusual. In these examples, the ACF values changed by an average of ~1.8% during the lifetime of the Scanning Electron Multiplier (SEM). Collectively, these observations suggest that more reliable ACF values can be determined by comparing pulse-mode and analog-mode readings separately for each element (rather than as an average for all elements) and within a session (rather than as a separate experiment).

**4.1 ExtractDat**

We have implemented a Python package called ExtractDat that decodes dat files as described above. As a package, ExtractDat provides classes for decoding the dat file header, masses, and scans. ExtractDat can also be run as an application, in which case it decodes the specified dat file(s) and produces a CSV containing the decoded information. ExtractDat is available on github at https://github.com/jhh67/extractdat.git.

**5.1 Conclusions**

This decoding routine extracts intensities from binary dat files that are generated during data acquisition with a Thermo Element$^{TM}$ ICP mass spectrometer. Comparison with values in FIN2

files suggests that most values extracted from dat files are identical to the values generated by the Thermo data reduction software.

Offsets of dat-file and FIN2-file values occur in both low-intensity and high-intensity signals. Offsets of low-intensity measurements occur due to omission of zero or low-intensity readings where multiple readings are acquired on each mass. These omissions result in significant errors in the average values reported in FIN2 files. Minor offsets also occur in a small percentage of high-intensity readings.

One of the benefits of having access to the measured signal intensities is the capability of calculating a separate ACF for each mass. As shown on Figure 3, average ACF values calculated for different elements differ by as much as 7.4%, and change by ~1.6%, during a typical 3.5-hour analytical session. As shown on Figure 4, such a change in ACF is typical for every analytical session, regardless of the condition of the SEM. Using decoded dat file intensities to determine an ACF for each mass, during an acquisition rather than as a separate experiment, is accordingly recommended.

An additional benefit of using the decoded data is that dat files are created by the data acquisition software immediately after each acquisition. In contrast, creating FIN2 files requires opening and operating the Result Display software, which adds a significant delay between analyses. Extraction of data from dat files accordingly allows for real-time data processing, which is advantageous because acquisition parameters can be optimized during an analytical session.

## Acknowledgements

## References

Chapman, J., Gehrels, G., Ducea, M., Giesler, D., Pullen, A., 2016. A new method for estimating parent rock trace element concentrations from zircon. Chemical Geology 439, 59-70.
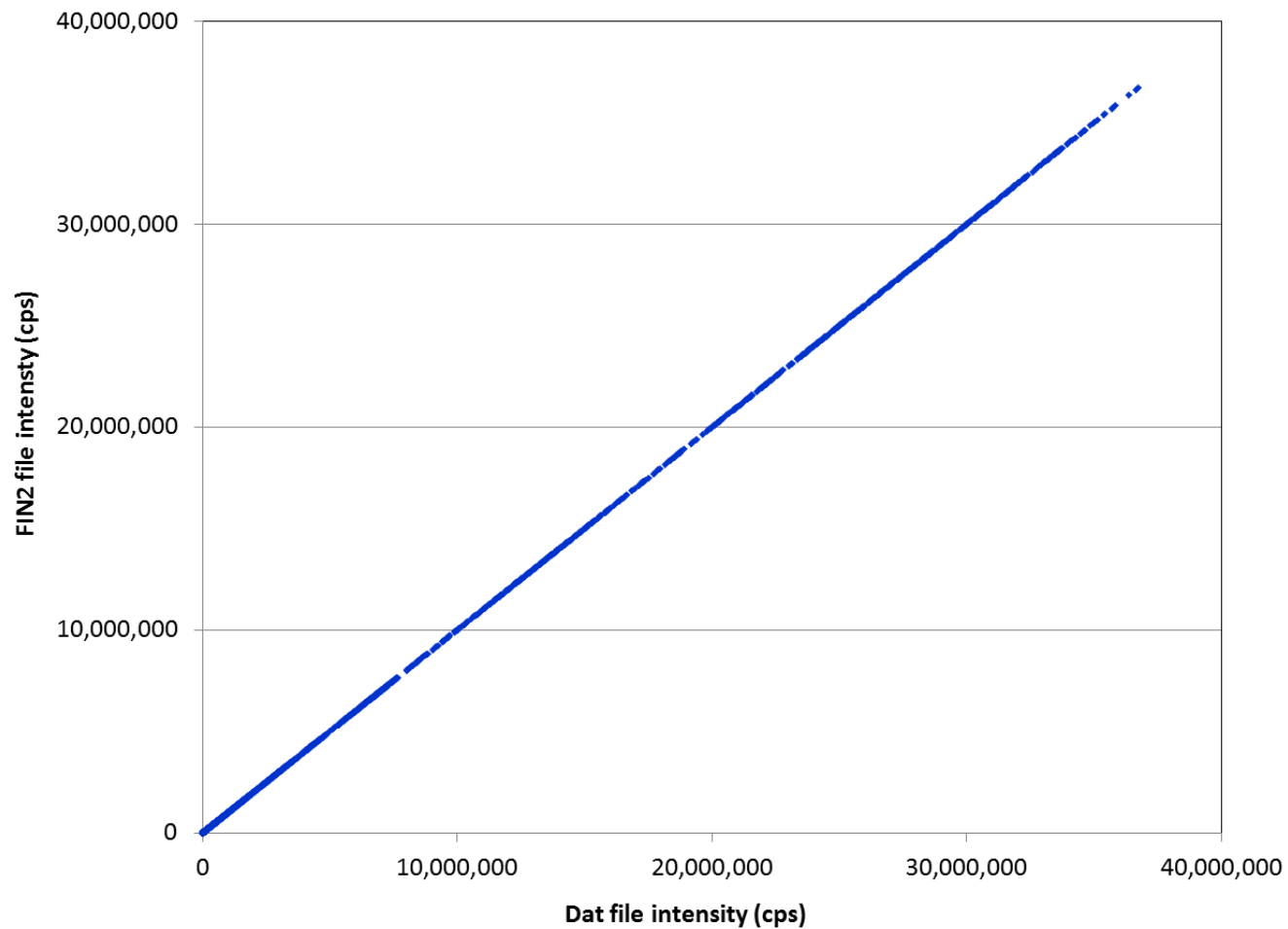
**Figure 1.** Comparison of FIN2 values with the average of measured intensities extracted from the decoded dat files. The linear relationship documents the excellent correspondence of most values.
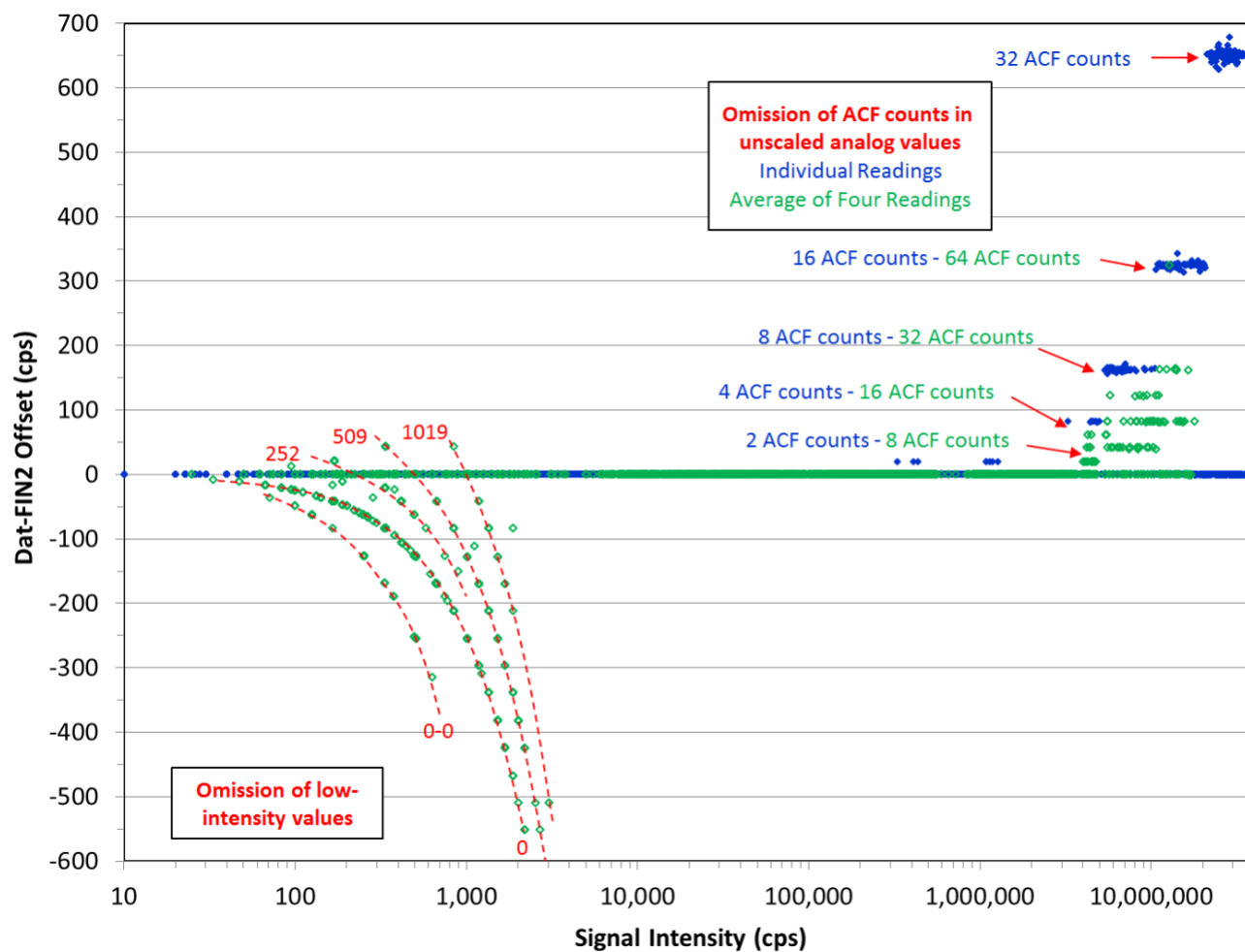
**Figure 2.** Deviation of FIN2 values from Dat-file values.
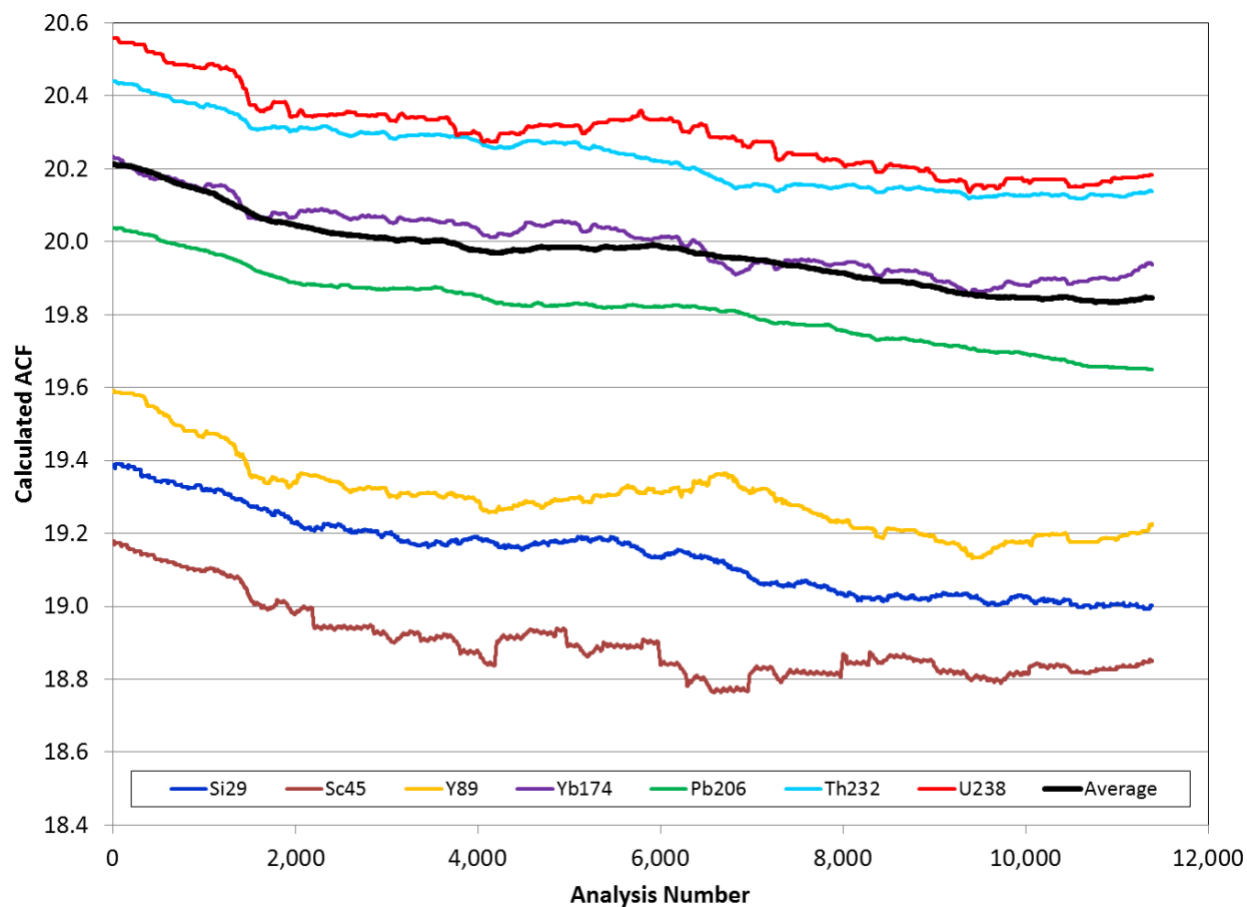
**Figure 3.** Value of ACF for each element during a single 3.5 hour analysis. Values shown are smoothed with a sliding window average of the closest 2,000 values. Data are available from Appendix 2.
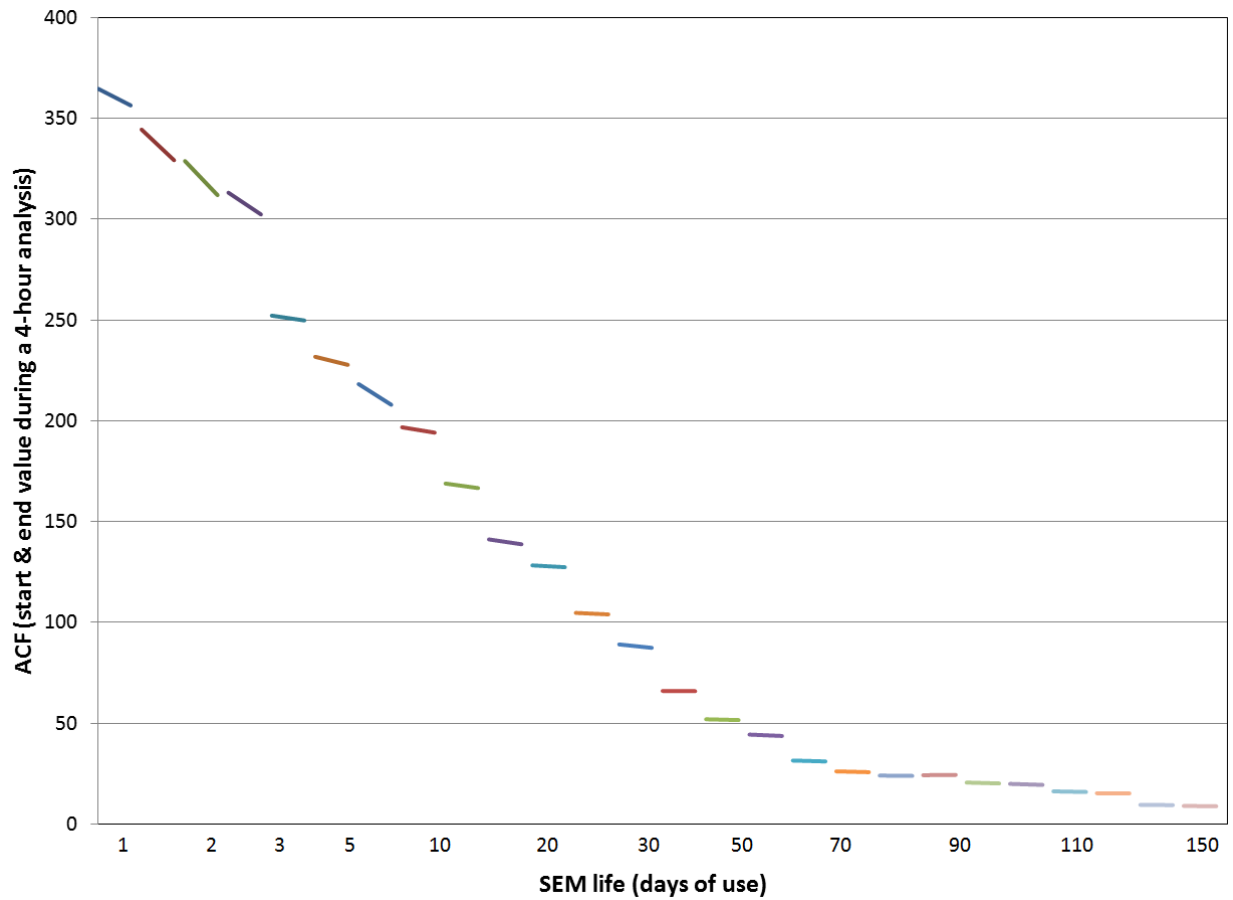
**Figure 4.** Average ACF values determined from 4-hour sessions during which the forty elements shown in Figures 1 and 3 were measured. Horizontal axis is the number of days since installation of a new SEM. Vertical axis values for each session are the initial and final values of the ACF as determined by a least-squares regression of measured ACF values. Data are available from Appendix 3.

**Appendix 1**: Excel file that contains an example of the output from our dat file decoder, and explains low-intensity and high-intensity errors. This appendix contains the data used to create Figures 1 and 2.

Sheets:

- **Header Info**: Explanation of information in each column. Information is for five analyses of 30 isotopes, with 73 rows per scan. Values shown are from decoded DAT files and from FIN2 files.
- **Low Intensity Errors**: Explanation of DAT and FIN2 values shown in various columns. Cells with low-intensity errors are highlighted in yellow.
- **High Intensity Errors**: Explanation of DAT and FIN2 values shown in various columns. Cells with high-intensity errors are highlighted in yellow.
- **FIN2-DAT plot**: Plot comparing DAT and FIN2 values for 10,950 measurements.
- **FIN2-DAT offset plot**: Plot comparing small offsets of DAT and FIN2 values for both low-intensity and high-intensity signals.
- **DAT-FIN data**: Data used for DAT-FIN2 plots.

**Appendix 2:** Excel file showing ACF values calculated for each element during a single 3.5 hour analytical session. This appendix contains the data used to create Figure 3.

Sheets:

- **ACFdata**: Analog and Pulse readings from 11,388 measurements (73 scans for 156 analyses) of six elements (29Si to 238U) during a single 3.5 hour session. ACF values are calculated for analyses with pulse-mode intensities less than 5M cps and greater than the user-selected minimum value (shown in cell I1). Values shown in the accompanying plot are sliding window averages of the closest 1000-2000 values.
- **ACFplot**: Plot of ACF values for each element during a single 3.5 hour session. Slider allows user to select the minimum intensity cut-off value for determination of ACF.

**Appendix 3:** Excel file showing how ACF value changes during four-hour analytical sessions through the lifetime of an SEM. This appendix contains the data used to create Figure 4.

Sheets:

- **Plot of ACF Value**: Plot showing ACF values for twenty-six four-hour sessions over ~150 days of use.
- **ACF Values**: ACF values determined from 28,981 integrations during each four hour session. Values at the bottom of each column are the initial and final values as

determined by a least-squares regression through all values. These initial and final values are shown in the companion plot.