SPUD EMTF Database: Change Log

Anna Kelbert, USGS

January 2, 2018

Here, we only document major changes to the database that all users should be aware of. New data submissions may appear at any time without additional documentation.

3–5 January 2018 – *Data Rotations*

Bug fixes

This release includes a major bug fix. In devising the XML schema, we have attempted to maintain compatibility with the historical Society of Exploration Geophysicists (SEG) Data Interchange Standard 1987, also known as EDI, to the extent possible. This includes maintaining consistency in the element names (e.g., Z, T, etc) and the statistical estimates, where relevant. However, the definition of **variances** in the EDI files, as we surmised after reading the manual extra carefully, deviates from the definition that we've been using for computation of variances from the covariance matrices in Z-files, by a factor of 2. Specifically, variances in EDI files are given for a complex number, not for the real or imaginary part, and are therefore twice those that we computed from Z-files. This inconsistency resulted in different definitions of Z.VAR and T.VAR between the data originally converted from Z-files, and data that came from the historic EDIs.

This update eliminates this inconsistency. From now on, all XML files include the variance as defined by the historic EDI format, i.e., the variance of the complex transfer function component. To compute the standard error of a real or imaginary part, e.g., for plotting, one would divide the variance by 2, then take square root (i.e. $Std = \sqrt{Var/2}$). To plot the error bars for a real or imaginary part, one would then multiply that number by 2.

We have now adjusted (multiplied) Z.VAR and T.VAR by the factor of 2, and re-archived **all data** that are affected by this change. Your work may have been affected if a) you used the error bars in the data set (e.g., for inversion) as recorded in the XML files or the EDI files produced by the database system, and b) you used these data in combination with historic surveys converted from EDIs.

Affected entities include all data that were originally created from Gary Egbert's format Z-files. These are: Argentina, CAFE-MT, Ethiopia, SAGE LP, USArray, USGS-GEOMAG, YSRP.

A second major bug fix involves correctly interpreting the rotations in EDI files. The old database had a warning to the effect that EDI files were not rotated. In the previous release of this database, we were trying to be careful about the error bars which can be corrupted if an EDI file is rotated in the absence of the full error covariance matrices. Unfortunately, this resulted in some occasions of file archiving where the data were not rotated to geographic, while the rotation information (ZROT) was not preserved in the archive. This is now corrected, and all data are now rotated to geographic with a caveat that the error bars are not to be trusted on these occasions.

This fix also involves the rotation of tipper data in EDI files. Specifically, in the case of an EDI file that only has ZROT information (i.e., impedance rotation) but no indication as to how the tipper are rotated (e.g., no TROT), the code defaulted to assuming that the tippers are oriented according to the channels orientations. However, it is much more typical for people who omit TROT information to mean that the tippers are oriented the same way as the impedance. So we have now changed the code's behavior to default to ZROT information for the tippers, if TROT or any other indication of tipper rotations is missing.

Affected entities: UofAdelaide, Lithoprobe Alberta Basement surveys. As far as we could discern, University of Alberta (UofA), GeoscienceAustralia, CAFE-MT, SAMTEX and SAGE data were provided in geographic coordinates and are not affected by these changes, but some additional SAMTEX and SAGE WB data has now been archived thanks to the improvements in the code that allow to archive EDI SPECTRA. Also note that CAFE-MT had been resubmitted to us last summer with rotation corrections applied.

Conceptual and schema changes

MT and other EM TF data are recorded in the field in a wide variety or orientations. Most often, the channels are oriented in an "X", "L" or "T" shape such that both HX and EX point to either geomagnetic or geographic coordinates. However, in certain challenging field circumstances that occur more often than one might think, the channels are oriented to some other, arbitrary, directions that possibly deviate from orthogonal.

For correct interpretation of the data, it is critical that these channel orientations are correctly recorded. However, in an analysis that wishes to use the measurements jointly with other data, it is also critical that the user rotates the transfer functions to a common orthogonal coordinate system. Usually scientists want the data to point to geographic North.

Many old data files are ambiguous with respect to data orientation, and often miss critical pieces of information, or contain several conflicting messages. Moreover, EDI data files (unless they include SPECTRA) do not contain sufficient information for statistically correct rotation of error bars. Our original intention was, therefore, to archive the data as they are, without a data rotation. We did not include a data rotation in the XML files, and presumed that the data are oriented to site layout, as defined by the input and output channel orientations, example:

```
<InputChannels ref="site" units="m">
<Magnetic name="Hx" orientation="0.0" x="0.0" y="0.0" z="0.0"/>
<Magnetic name="Hy" orientation="90.0" x="0.0" y="0.0" z="0.0"/>
</InputChannels>
<OutputChannels ref="site" units="m">
<Magnetic name="Hz" orientation="0.0" x="0.0" y="0.0" z="0.0"/>
<Electric name="Ex" orientation="0.0" x="-50.0" y="0.0" z="0.0" x2="50.0" y2="0.0" z2="0.0"/>
<Electric name="Ey" orientation="90.0" x="0.0" y="-40.0" z="0.0" x2="0.0" y2="40.0" z2="0.0"/>
</OutputChannels>
```

We have since come to realize that this strategy can result in severe data misinterpretation by scientists who are not well versed with certain subtleties of MT data. Moreover, since the channel orientations in the file had to match the orientation of the TFs, it didn't allow us to preserve the original site layout, making any rotation irreversible. Finally, a scientist couldn't quickly discern whether or not the data were oriented to orthogonal geographic or to some other layout, without checking the orientations of all channels. Even though we try to specify the declination, if known, there wasn't enough information in the files to know whether or not the data were collected in geomagnetic coordinates in the first place.

We have decided to adjust the XML schema to include **unambiguous rotation information**. Our new strategy is to always rotate the transfer functions to orthogonal geographic for archival at the database. This is a convention that we have adopted for user convenience. We should note that the updated XML format has enough information to unambiguously store data in any orientation, and also allows for reversible rotations. For this, we had to redefine the meaning of the channel orientations completely.

First of all, we added a new element to the XML `<Site>` header, right after `<Location>` (it's that important!). Two options are allowed:

```
<Orientation angle_to_geographic_north="0.0">orthogonal</Orientation>
```

or,

```
<Orientation>sitelayout</Orientation>
```

In the database, the files will be oriented to orthogonal geographic, as indicated by the first variant of the new `<Orientation>` element. Both `orthogonal` and `sitelayout` are keywords, and no other keywords are supported. More generally, the angle to geographic North can of course be arbitrary. Alternatively, if the orientation is defined by the site layout, it no longer necessarily has to be (or presumed to be) orthogonal. Even as the data are archived in orthogonal geographic, we will always strive to archive the channel information in their original site layout. If the data are oriented to site layout, then and only then the channel orientations will also define the data orientations. Codes have been developed to revert any rotation back to the site layout, if needed, as well as to rotate to any other arbitrary orthogonal coordinate system. These codes will be made freely available.

To better match the new meaning of the channels, we have now encompassed them with a new element, `<SiteLayout>`. The channels are no longer rotated with the data, as they merely indicate the original site layout, whenever this information is known. No other changes to the channel information have been made in this update. Example:

```
<SiteLayout>
<InputChannels ref="site" units="m">
<Magnetic name="Hx" orientation="0.0" x="0.0" y="0.0" z="0.0"/>
<Magnetic name="Hy" orientation="90.0" x="0.0" y="0.0" z="0.0"/>
</InputChannels>
<OutputChannels ref="site" units="m">
<Magnetic name="Hz" orientation="0.0" x="0.0" y="0.0" z="0.0"/>
<Electric name="Ex" orientation="0.0" x="-50.0" y="0.0" z="0.0" x2="50.0" y2="0.0" z2="0.0"/>
<Electric name="Ey" orientation="90.0" x="0.0" y="-40.0" z="0.0" x2="0.0" y2="40.0" z2="0.0"/>
</OutputChannels>
</SiteLayout>
```

Additional information (such as "geomagnetic") may be specified as a `SiteLayout` attribute, but we opted against providing this information. It can be discerned from the channels orientations, which provide a significantly more general way to record this information. Duplication of information in both verbal and numeric form leads to conflicts and confusion more than to clarity.

Finally, we have also added an optional descriptive element called `<RotationInfo>`. It has helped us to record any subtleties and ambiguities with respect to discerning the data rotation that we had to overcome before the data could be archived. Our human interpretation of human omissions from 20-30 yrs ago is not flawless. We consider these considerations an important piece of information for anybody who has concerns about any particular data site and needs to find the root of the problem.

Code improvements

The updated EMTF FCU v4.0 code includes numerous improvements to correctly parse and rotate many variants of EDI files. Among the major improvements are the ability to correctly read EDI SPECTRA files and compute the full error covariances from such files. This allows for arbitrary rotations of the transfer functions without loss of information, even after conversion to the XML. We have also now implemented frequency-by-frequency rotations that allow conversion to XML of data that went through a principal axis rotation algorithm. This can now be undone for EDI files that include variable ZROT values, even though the resultant error estimates may suffer. Finally, we have implemented a general rotation algorithm that allows for rotation from an arbitrary (not necessarily orthogonal) set of channel orientations, to an arbitrary orthogonal coordinate system, or back. All of these improvements have allowed us to archive a large number of historic MT TFs that were previously unaccessible for archiving.

Limitations / To do

XML files include links to **HTML definitions** of included data types, as well as all pertinent statistical estimates. These HTML files are yet to be completed and uploaded. In the meantime, please note that these definitions are consistent with the EDI manual.

Derived data types are supported by the XML format. However, correct rotation of derived data types requires recomputation of the primary data types (e.g., impedance, tipper), their rotation, followed by a recomputation of the derived data (e.g., apparent resistivities and phases, tipper magnitude, skew, phase or ellipticity). General implementation of such a capability is a challenging task that is not currently warranted by data archiving needs. Therefore, for now we omit any additional derived data type products on conversion to XML files. If you have a need in any such products, the original EDI files are always included in the archive bundle.

Similarly, any general statistical estimate is supported by the XML format, as long as that estimate relates one or more of primary data type components to each other. However, the inhomogeneity with which, e.g., the **coherence and predicted coherence** information are sporadically recorded in the EDI files, is such that correctly reading and interpreting this information in general is challenging. This work is currently not warranted by the few occurrences of these estimates in the historic EDI files, and the coherences are therefore omitted from the XML files. As with the derived data types, they can always be accessed by downloading the original EDI file.

Finally, during the archiving work we have encountered several occasions of EDI SPECTRA files that have been edited, perhaps manually, to include fewer frequencies than the **NFREQ** value might suggest. Since we want the output XML file to contain a correct frequency count, we have not attempted to overcome this problem programmatically (although we may implement a workaround in the future). In these rare circumstances, conversion to XML will halt and the original EDI needs to be edited to fix the problem (by adjusting NFREQ value to the actual number of frequencies in file).

26 February 2016 – *Data Citations*

Added **Digital Object Identifiers (DOIs)** for Electromagnetic Transfer Functions. DOIs, attributed directly to data sets, allow us to give proper credit to the very significant effort of data collection, allowing to cite the data sets directly in any new publications in lieu (or in addition to) data interpretation papers. This provides a great incentive of data sharing that was never before in place in magnetotelluric community. DOI attribution has additional benefits, which include the ability to track the usage of any particular data set and notify the users of any updates as necessary.

To make this possible, we reached out to the authors of every data set that we archived or were planning to archive to obtain the critical information for the data citation, namely the authors, years of data collection, title, acknowledgements, and selected publications. We found that this is a learning process for everyone involved. The concept of data citation was so new to this community that most did not understand our questions at first, thinking that its a publication reference that we were after. When we explained that this was truly a way to give credit for data collection (NOT for data interpretation as has been customary until now), many of our colleagues came to value and appreciate this opportunity. The authorship of the data sets often turns to be notably different from that in the final publications, if such publications exist.

The greatest technical challenge was achieving the right granularity of the DOIs. A unique DOI is given and automatically included within any XML data file upon submission to EMTF database. However, for data citations to be practical in publications, one would need a data DOI that points to a collection of data, such as a survey, united by the common authors, dates, geographical area, and purpose. Jointly with IRIS, we have devised a strategy to attribute a unique DOI to a geophysical survey. This goes under a "Survey DOI" and is available along with the complete data citation, whenever any site is selected and opened in the database. Users are strongly encouraged to cite Survey DOIs in their publications, just like they would cite a paper. All data in the database were re-archived with the authorship attribution information.

For example, to cite the USArray magnetotelluric transfer functions, one would use the following reference:

Schultz, A., G. D. Egbert, A. Kelbert, T. Peery, V. Clote, B. Fry, S. Erofeeva and staff of the National Geoelectromagnetic Facility and their contractors (2006-2018). "USArray TA Magnetotelluric Transfer Functions". doi:10.17611/DP/EMTF/USARRAY/TA. Retrieved from the IRIS database on Feb 26, 2016

The date of retrieval from the database would reflect the date when the data were accessed. This citation should appear in the References section of the manuscript. In the text of a paper, this citation looks like "Schultz *et al* (2006-2018)". Some journals complain when they see this at first, but they all are eventually ok with this format.

Data acknowledgements are, ideally, included verbatim in the relevant section of the paper. For example, here's the full acknowledgement for USArray:

USArray MT TA project was led by PI Adam Schultz and Gary Egbert. They would like to thank the Oregon State University MT team and their contractors, lab and field personnel over the years for assistance with data collection, quality control, processing and archiving. They also thank numerous districts of the U.S. Forest Service, Bureau of Land Management, the U.S. National Parks, the collected State land offices, and the many private landowners who permitted access to acquire the MT TA data. USArray TA was funded through NSF grants EAR-0323311, IRIS Subaward 478 and 489 under NSF Cooperative Agreement EAR-0350030 and EAR-0323309, IRIS Subaward 75-MT under NSF Cooperative Agreement EAR-0733069 under CFDA No. 47.050, and IRIS Subaward 05-OSU-SAGE under NSF Cooperative Agreement EAR-1261681.

If you heavily use USArray MT data in your publication, you are encouraged to include the complete text of this acknowledgement in your own acknowledgement for the paper. If the word limit doesn't allow that, please truncate it as necessary but still include it. Please note that a mere acknowledgement of the data, without a data citation as provided above, is not appropriate. This data citation is the only credit these people get for their many years of hard work in data collection and processing. Thank you to those of you who already follow these citation practices, that's much appreciated.

To cite the EMTF database, please use:

Kelbert, A., G.D. Egbert and A. Schultz (2011), IRIS DMC Data Services Products: EMTF, The Magnetotelluric Transfer Functions, https://doi.org/10.17611/DP/EMTF.1.

This information is also accessible by clicking the "Citations" link at the top of EMTF database website.