archaeometry

Archaeometry **59**, 4 (2017) 714–730 doi: 10.1111/arcm.12282

CHARMED PYMCA, PART I: A PROTOCOL FOR IMPROVED INTER-LABORATORY REPRODUCIBILITY IN THE QUANTITATIVE ED-XRF ANALYSIS OF COPPER ALLOYS*

A. HEGINBOTHAM

J. Paul Getty Museum, 1200 Getty Center Drive, Los Angeles, CA 90291, USA and Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands

and V. A. SOLÉ

European Synchrotron Radiation Facility, CS 40220 38043 Grenoble Cedex 9, France

This paper describes a protocol for quantification of heritage copper alloys by energy-dispersive X-ray fluorescence spectroscopy (ED-XRF). The protocol, nicknamed CHARMed PyMca, is designed for users who wish to maximize inter-laboratory reproducibility of quantitative ED-XRF results for the wide range of copper alloys found in heritage materials. By maximizing reproducibility, this protocol should facilitate collaboration and allow the rigorous use of shared data and databases. The protocol uses free, open-source, fundamental parameters software called PyMca. PyMca allows for a consistent and transparent application of the fundamental parameters approach independent of the ED-XRF instrumentation used. The proposed protocol calls for calibration of standardless PyMca results against a set of certified reference materials designed specifically for use with heritage copper alloys, the so-called copper CHARM set. Finally, this protocol calls for the calibration-to-standards to be carried out following a consistent strategy, including error modelling and the incorporation of a validation procedure. A reproducibility study was conducted using CHARMed PyMca and eight different ED-XRF instruments of six different types. In comparison to a 2010 study conducted according to the same method, CHARMed PyMca showed a dramatic improvement in reproducibility and method sensitivity.

KEYWORDS: XRF, CALIBRATION, COPPER, REPRODUCIBILITY, FUNDAMENTAL PARAMETERS

INTRODUCTION

The study of heritage copper alloys (HCAs) is in a very dynamic period and the volume of quantitative compositional data being produced is growing very rapidly. These data are being used to deduce diachronic and geographical trends useful for reconstructing technological evolution, trade in materials and provenance. In particular, the use of energy-dispersive X-ray fluorescence spectroscopy (ED-XRF) to study copper alloys has continued to expand dramatically due to the ever-decreasing cost and increasing portability and ease of use of ED-XRF instrumentation. With this growth comes an increasing interest in aggregating existing quantitative XRF data on HCAs in order to both broaden and deepen our insights through collaboration and meta-studies (Frank and Pernicka 2012; Bray *et al.* 2016). Rehren and Freestone (2015), writing about a parallel

*Received 19 September 2015; accepted 29 July 2016
†Corresponding author: email aheginbotham@getty.edu

evolution in the compositional study of glass, state that progress towards deeper understanding (asking 'why' and 'how' rather than just 'what', 'where' and 'when') relies on 'expansion and refinement of the [shared] data base'. It seems evident that the study of copper alloys would also benefit tremendously from the continued growth of a shared body of quantitative compositional data (Rehren and Freestone 2015).

There is, however, an elephant in the room. If one wishes to analyse a database of research results that combine quantitative compositional data from different collaborating laboratories, good (or at least, *known*) inter-laboratory reproducibility is a fundamental prerequisite. This goal has not been simple to achieve for copper alloys in the art and archaeology domain (Heginbotham *et al.* 2011). Inter-laboratory reproducibility of compositional analysis is increasingly recognized as a potential barrier to effective collaboration in the realm of cultural heritage, not just for copper alloys, but also with regard to other materials (Frahm 2013; Speakman and Shackley 2013; Rehren and Freestone 2015).

Unfortunately, the pursuit of better reproducibility in the quantitative study of heritage copper alloys with ED-XRF faces some unique challenges. Even under the best of conditions (i.e., with an uncorroded, flat surface and a homogeneous matrix that is representative of the bulk) the large number of elemental analytes encountered, the large concentration ranges expected for each and the significant variability in overall matrix characteristics combine to create an environment that is extremely challenging for quantitative ED-XRF.

This paper presents a protocol, nicknamed 'CHARMed PyMca' for the quantification of heritage copper alloys by ED-XRF that is designed specifically to maximize inter-laboratory reproducibility for the wide range of alloy types found in heritage materials. The name refers to the two essential components of the protocol, namely PyMca fundamental parameters software, used in conjunction with the so-called copper CHARM (Cultural Heritage Alloy Reference Material) set of certified reference standards. The protocol assumes that the metal to be analysed is in a state appropriate for study by ED-XRF (uncorroded, flat, homogeneous and representative). Under such circumstances, it is designed to provide accurate results with well-characterized precision, and to do so for a broad range of elements over a large concentration range, independent of instrument type.

The protocol calls for the use of a free, open-source, fundamental parameters software for spectral analysis called PyMca (Solé *et al.* 2007). The fundamental parameters approach was selected for this protocol because it is generally favoured over empirical methods for quantification where significant matrix variability may be present and where large concentration ranges are to be addressed (Lachance and Claisse 1995, 356–7; de Vries and Vrebos 2002; Mantler *et al.* 2006). The fundamental parameters approach has also recently been shown to be strongly associated with good inter-laboratory reproducibility in practice (Heginbotham *et al.* 2011). PyMca software was selected for the protocol primarily because it is available to any interested user at no cost, and it allows for a consistent and transparent application of the fundamental parameters approach independent of the ED-XRF instrumentation used.

The CHARMed PyMca protocol calls for the calibration of standardless PyMca results using a specific, widely available set of high-precision certified reference materials designed specifically for use with heritage copper alloys, the so-called copper CHARM set (Heginbotham *et al.* 2015). The use of this rigorously designed and fabricated set ensures that the user's results will be as accurate as possible, and valid over as large a concentration range as possible. Finally, this protocol calls for the calibration-to-standards to be carried out following a consistent strategy, including error modelling and the incorporation of a validation procedure.

The CHARMed PyMca protocol has been evaluated according to the same methodology as the 2011 reproducibility study mentioned above (Heginbotham *et al.* 2011). The results of this new

study, using eight different instruments, demonstrate that the use of the CHARMed PyMca protocol yields a significant improvement in reproducibility, accuracy and method sensitivity. A brief summary of these results is given below, and a thorough discussion and interpretation will follow in part 2 of this paper.

All aspects of the protocol are intended to be fully transparent, and to provide the individual user with results that are accurate and, perhaps more importantly, *well-characterized* in terms of precision. By pursuing these goals, the protocol intends to provide *collaborating* users with a means towards significantly improved inter-laboratory reproducibility.

DESCRIPTION OF THE PROTOCOL

The CHARMed PyMca protocol for quantitative analysis of heritage copper alloys has been developed and tested using several tube-based ED-XRF instruments from different manufacturers, typically operated between 40 and 50 kV with moderate filtration in an air path environment (although a vacuum or helium flush protocol could also be used). Several tube anode materials have been used (Cr, Re and Rh) and both PIN and silicon drift detectors have been employed. The protocol is presented below in five sections: the use of PyMca software; the use of the copper CHARM standard set; the calibration strategy; the error modelling strategy; and the validation procedure.

PyMca

A fundamental parameters (FP) approach to spectral analysis is proposed here based on its suitability to the copper alloy environment, where matrix variability is great and concentration ranges are large (Lachance and Claisse 1995; de Vries and Vrebos 2002; Mantler *et al.* 2006). In the early period of XRF development, the use of the FP approach was restricted due to the considerable computational power required. Partly as a result of restricted access to advanced computational facilities, a wide array of quantitative techniques that required less computation were developed to account for absorption and enhancement effects that confound accurate XRF analysis (so-called matrix effects). These techniques generally involve the use of influence coefficients, either derived empirically or generated theoretically. Today, however, the computational power required to apply fundamental parameters to ED-XRF analysis is readily available to any analyst at relatively low cost and the use of the robust FP approach can be readily adopted.

In addition to the theoretical arguments for the use of fundamental parameters, a round-robin inter-laboratory study conducted in 2009–10 (Heginbotham *et al.* 2011) suggests that FP methodology, combined with the use of reference standards, appears to deliver superior reproducibility compared to empirical methods in the analysis of copper alloys. This enhanced reproducibility was observed even though the six laboratories using 'FP with standards' methodology used six different instrument types and six different FP software packages. Given this observation, one might well ask why advocate for the use of PyMca software specifically when any FP software might perform as well? There are a number of arguments to be made in favour of PyMca, which can be condensed into three classes: accessibility, transparency and functionality.

In terms of accessibility, PyMca is a freely downloadable, open-source software that can be used on Solaris, Linux, Windows and Mac OS X platforms. The software is institutionally supported by the European Synchrotron Radiation Facility (ESRF) and is maintained and updated frequently. The software is able to process spectra from many instrument types; to date, spectra have been processed from Bruker's Artax and Tracer spectrometers, Thermo's Niton and ARL

spectrometers, Olympus' Delta spectrometers and XGLab's Elio spectrometers. Additional spectrum formats can be quickly made readable by request to the program's administrator.

Transparency is an extremely important factor that also weighs in favour of PyMca. Most proprietary quantification software packages (FP or otherwise) do not allow the user access to, much less full control over, the many parameters that must be configured for a successful quantification program. This can result in situations where different software versions produce systematically different results on the same samples and the user cannot explain or correct the differences by making adjustments to the software configuration (Goodale *et al.* 2012). Even in instances where the user is able to customize calibration routines, if the software is proprietary, sometimes 'full disclosure and discussion of the calibration routine is not possible' (Rowe *et al.* 2012).

PyMca, in contrast, strives to be completely transparent in its methodology and allows the user full control over virtually all parameters affecting its standardless quantification process. PyMca offers full control over background/continuum modelling, peak shape modelling (including long, short and step tailing), element line group modelling, energy calibration, and pile-up and escape peak modelling. The software also gives the user control over the modelling of energy output from polychromatic X-ray sources, which is very important for most investigators in art and archaeology, who predominantly use X-ray tube-based XRF instruments (Solé *et al.* 2007).

Recent improvements in PyMca have had the effect of greatly improving its performance in the quantitative analysis of complex high-density materials such as copper alloys. In particular, starting with version 5.0, PyMca models both secondary and tertiary excitation phenomena for every assigned peak in the spectrum. The modelling strategy is based on the work of D. K. G. de Boer; for details of the implementation, see de Boer (1990) and Solé *et al.* (in preparation). A second, major improvement in PyMca, beginning with version 5.0, is the implementation of reiterative matrix modelling in which the initial matrix composition can be automatically refined. In the case of copper alloys, the initial composition can be as simple as pure Cu, but PyMca's initial estimate of the composition will then be used as the matrix description for a second iteration of the quantification process. The second estimate of composition can then be used for a third iteration and so on. The number of iterations and the specific elements to be considered in the matrix and their chemical form are user-configurable.

For all of the sophisticated capabilities that PyMca has for the analysis of XRF spectra, quantitative results generated by the software are still prone to some degree of systematic error. The likely causes of such error fall into three general categories.

Theoretical Uncertainties

The theoretical database used by PyMca for calculations contains thousands of values, drawn from the published literature, for constants such as binding energies, fluorescence and Coster–Kronig yields, photoelectric absorption cross-sections, radiative emission probabilities and mass attenuation coefficients. These values have some uncertainty associated with them and this is likely to lead to some error in the spectrum modelling (Caussin 2013; Schoonjans *et al.* 2013). In addition, PyMca does not account for fluorescence induced by ejected photoelectrons or Auger electrons within the matrix, which may be significant, particularly for low atomic number elements and with higher energy excitation (de Vries and Vrebos 2002; Fernandez *et al.* 2013, 2014). Perhaps the most important theoretical uncertainty, however, comes with the spectral modelling of X-ray tube output. PyMca relies on the formulae reviewed and outlined by Ebel (1999) to generate a spectral model of the radiation emitted by X-ray tubes. While this approach yields a reasonable approximation of reality, there is clearly a significant degree of uncertainty in

the models and this can result in biased estimates of concentration for certain elements depending on their binding energies.

System Description

At least as important as the problems associated with theoretical uncertainties is the fact that PyMca relies on an accurate description of the total instrument and sample system in order to make accurate quantitative calculations. The list of system characteristics that must be included in PyMca's 'configuration file' for any specific XRF instrument is compendious and includes such information as the tube anode material, the tube voltage, the current, the analysis time, the geometry of the system (incident X-rays, sample and detector), the angle of electron incidence on the anode, the angle of X-ray emission from the anode towards the sample, the scattering angle, attenuators (including tube filters, tube and detector windows, air path and so on), the detector thickness, the detector response function, sample homogeneity and a list of possible elements present in the sample. In reality, perfectly accurate system modelling is difficult, if not impossible, to achieve and so some degree of uncertainty and bias in PyMca's standardless results may be expected.

Deconvolution Difficulties

Despite PyMca's sophistication and customizability, deconvolution of overlapping peaks can still pose problems in certain circumstances, such as extracting a small peak from the tail of a much larger peak, as is often the case with a small Ni–K_α peak on the tail of a dominant Cu–K_α peak. In instances such as this, consistent deconvolution may be difficult to achieve given the wide variety of matrix types encountered in HCAs, and this may affect PyMca's ability to produce consistent and accurate results.

CHARM

In order to compensate for potential errors in PyMca quantification, CHARMed PyMca calls for calibrating and correcting the standardless results generated by PyMca using the copper CHARM set of 12 certified reference materials (in practice, we have also used the two supplementary high-arsenic standards, bringing the total number of standards used in the examples presented to 14). The design advantages of the set have been previously published in detail (Heginbotham *et al.* 2015). Some of the most important advantages are that the copper CHARM set provides a common reference set with a very broad concentration range for 20 elements, 15 of which are regularly found in air-path XRF analysis of heritage copper alloys. The standard set includes a wide variety of alloy types and intentionally varied element ratios designed to challenge any quantification methodology, and the high precision of the certified values in the set (typically ±1–2% of the certified value) allows the uncertainties in these values to be disregarded during regression analysis, greatly simplifying the calibration and error modelling equations. Finally, the set should be available to any interested researchers for many years to come through the manufacturer, MBH Analytical Ltd (http://mbh.co.uk/).

In this proposed protocol, the standards are to be analysed three times each, on different days, preferably by different operators, in different locations, after a suitable warm-up of the X-ray tube. In this way, some measure of the instrumental variability can be incorporated into the calibration procedure. While more than three replicate measurements would clearly yield a more accurate estimate of the instrumental, or intra-laboratory, reproducibility of results, we have

settled on triplicate measurements in the interest of keeping the burden of the protocol reasonable. Furthermore, the triplicate measurements will be carried out on the entire calibration set of 12 standards for each element, which should provide adequate data to construct a useful error model (see 'Instrumental reproducibility' below).

Once the standards have been analysed, the resulting spectra can be processed with PyMca using the 'batch process' function to give initial standardless quantitative results that can be saved for export in several formats. Processing the standard spectra requires the creation of an optimized PyMca configuration file for the instrument being used. This can be a time-consuming process; however, once done for a specific instrument type, the configuration can easily be shared and used on similar instruments. Example configuration files that have been used by the authors for several different instruments can be downloaded at http://www.getty.edu/museum/conservation/papers/pymca.html.

Multi-Element Calibration

Once standardless PyMca results have been generated, a multi-element calibration can be built. To facilitate data sharing between laboratories, it is important to follow a rigorous and consistent calibration strategy. One weakness noted in the 2008 reproducibility study (Heginbotham *et al.* 2011) was an inconsistency in the specific elements analysed. Based on the shared experience of the authors of the round-robin study and the CHARM set publication, the authors recommend that for air-path protocols, individual calibrations be built for a minimum of 15 significant elements including Cr, Mn, Fe, Co, Ni, Cu, Zn, As, Se, Ag, Cd, Sn, Sb, Pb and Bi. The individual calibrations should be built following a best practices approach to both calibration procedure and error modelling based on accepted standard methodologies (Currie 1999; Burgess 2000; Barwick 2003; Institute for Reference Materials and Measurements 2010).

Fundamentals

There are several important features of a rigorous calibration scheme. First, the certified weight per cent (wt%) concentration of the standard set should be considered as the independent variable (X), whereas the PyMca calculated wt% concentration should be considered as the dependent variable (Y). Second, replicate measurements of the standards should be averaged before performing regression analysis on the data. This approach is significant and will change the results of the analysis because the true number of independent variables is equal to the number of standards, not to the number of measurements made. Variability within the replicate measures (instrumental reproducibility) is accounted for in a later step. An ordinary least squares linear regression can then be calculated to characterize the relationship between the certified and calculated values. The regression will yield an equation of the form $Y = aX + b$ and, subsequently, a newly acquired standardless PyMca quantitative result (y_i) can be used to predict the true composition of an unknown sample (x_i) by inverting the equation as $x_i = (y_i - b)/a$.

By default, it is recommended that non-normalized PyMca results be used to generate the regression equations. However, if the use of normalized data offers a significant advantage, the user may then choose to switch to this mode for the calibration of any specific element. There is no imperative to treat all elemental calibrations in the same manner. For each element, the analyst can evaluate the difference in goodness-of-fit between the two approaches by noting the different R^2 values associated with each.

A complete regression analysis produces a range of descriptive statistics and these should be carefully inspected. In particular, residuals plots are very useful for detecting evidence of non-linearity and heteroscedasticity in the data set. These plot the calibration residuals (the difference between the PyMca observed result and the predicted result based on the regression equation) against the true concentration for each standard. In the event that the residuals plot points strongly towards non-linearity for a specific element, a second-order regression may be merited for its calibration, particularly if it results in a significant reduction in residual standard deviation (RSD) compared to the linear model. In principle, such non-linearity should be very rare, although in practice the authors have noted occasional instances where calibrations based on high-energy lines (e.g., Ag, Sn and Sb) seem to benefit from a second-order model. If heteroscedasticity is suggested by the residuals plot, this should be taken into consideration when determining the error model to be used (see discussion below).

Matrix Correction

In the context of the analysis of complex copper alloys using the CHARM set, it is recommended that the regression models be inspected for any inter-element (matrix) effects apparent in the data for which PyMca has not been able to correct sufficiently well. This may be done by building a table of correlations between the residuals and the measured concentrations of other elements in each of the CHARM set standards in order to draw attention to elements that may be responsible for systematic errors in the PyMca results. Where high correlations exist, a graphical representation of the relationship (matrix element measured result versus residual) may help to determine whether or not the correlation might reasonably be interpreted to imply causation. The investigator must draw on his or her sound understanding of the fundamentals of X-ray fluorescence spectroscopy to determine whether the effect is plausible (based on X-ray physics or potential problems in the deconvolution of overlapping peaks). If the evidence clearly suggests that an uncorrected matrix effect is present, it should be permissible to generate and use a matrix correction factor (F_c) that can be used to adjust the standardless PyMca result (y_i) by a fixed fraction of the PyMca result for the selected matrix element (m_i) according to the formula $y_m = y_i - (F_c m_i)$, where y_m equals the matrix-adjusted PyMca result. The value of F_c can be determined iteratively such that the sum of squared residuals (SSR) of the matrix corrected regression is minimized. This correction factor can then be used in the application of the calibration to new experimental results. It should be stressed that PyMca should be expected to model and account for the vast majority of matrix effects prior to calibration, and thus the use of matrix correction in the calibration process should be very unusual and undertaken with caution. The limited number of calibration standards in the CHARMed PyMca scheme creates a danger of inappropriate 'overfitting' based on random correlations. The investigator should take a conservative approach and only apply a matrix correction if the correlation is very strong, the 'residuals versus matrix element concentration' plot shows a clear trend and the matrix effect is reasonable in a physical sense. The danger of overfitting can be further controlled by limiting matrix correction possibilities to one element and including the analysis of a validation set in the procedure (see below).

Forced Intercept

The initial regression model should also be inspected to confirm that the y intercept has been determined appropriately. In this context, the y intercept itself represents the most likely value that will be returned by PyMca if a sample is analysed the true composition of which for a given

element is zero. In some instances, calibration points in the high end of the calibration range may, by virtue of their high leverage, force the intercept of the preliminary OLS regression away from what would otherwise be the optimum line for the group of calibration points nearest to zero. This condition, if it exists, should be evident from looking at the residuals plot. Since even small absolute errors associated with small (i.e., trace) results can result in disproportionately large relative errors, the user may choose to force the intercept of the regression line in order to maximize accuracy for the calibration values nearest to zero. In the CHARMed PyMca environment, where the distribution of calibration values is weighted towards zero, this may conveniently be done by iteratively applying a forced-intercept linear regression to the calibration data and finding the value of forced intercept (and corresponding slope) that minimizes the SSR for the five or six lowest calibration points only. Before adopting a forced intercept calibration, it is beneficial to plot the calibration data with the forced regression line overlaying the default line so that the user can visually assess the merits of introducing a forced intercept.

FINAL CALIBRATION

Once the issues of normalization, residual matrix effects, linearity and intercepts have been addressed, a final calibration equation can be formulated for each of the 15 core elements. These equations can then be used to convert raw PyMca results to calibrated results. Calibrated results should not be normalized to 100%, as this would needlessly, and invariably, shift the results *away* from the best estimate of the true composition.

Error Modelling

Another extremely important aspect of building a multi-element calibration is to construct a useful error model. Error modelling is conceptually the most challenging aspect of the calibration procedure. In this area, it is perhaps best to take to heart Box and Draper's famous truism that 'All models are wrong, but some are useful' (Box and Draper 1987). In terms of usefulness, the authors bear in mind that the purpose of the entire CHARMED PyMca exercise is to develop a protocol that facilitates data sharing by maximizing inter-laboratory reproducibility. The authors thus propose that an error model be commonly adopted based on what Lloyd Currie calls the 'complete error budget' of the system. Currie argues that if a complete error budget is successfully modelled, the apparent dichotomy between intra- and inter-laboratory approaches 'essentially vanishes' (Currie 1999). In terms of rigour, this is easier said than done. It is clear from the authors' experience calibrating several types of XRF instrumentation that the dominant uncertainties associated with the CHARMed PyMca quantitative protocol are those associated with the calibration itself. Error associated with instrumental reproducibility is the second most important category of error but this is generally much smaller and often negligible. A useful estimate of the 'complete error budget' for each element in the calibration can then be constructed by taking these two error sources into account.

Error of Prediction

To estimate the uncertainty of a predicted value of an unknown sample based on a linear regression calibration, the accepted method is to calculate the standard error of prediction using the formula (given here as presented in Barwick 2003):

$$s_{xo} = \frac{s(r)}{m}\sqrt{\frac{1}{N}+\frac{1}{n}+\frac{\left(\bar{y}_o-\bar{y}\right)^2}{m^2\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}},$$

where n is the number of paired calibration points (x_i, y), m is the calculated best-fit slope of the calibration curve, N is the number of repeat measurements made on the unknown sample, $\bar{y}_o$ is the mean of N repeat measurements of y for the sample, $\bar{y}$ is the mean of the y values for the calibration standards, x_i is a value on the x-axis, $\bar{x}$ is the mean of the x_i values, and

$$s(r) = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}},$$

in which y_i is the observed value of y for a given value of x_i, and $\hat{y}_i$ is the value of y predicted by the equation of the calibration line for a given value of x_i.

A confidence interval is obtained by multiplying s_{xo} by the two-tailed Student's t value for the appropriate level of confidence and $n - 2$ degrees of freedom. A full discussion of the practical application of this formula is given in Burgess (2000) and Barwick (2003).

The error of prediction defines the range of concentrations that the true concentration of a test sample is likely to fall within, given a specified degree of confidence (commonly 95%). This is distinct from the more commonly cited standard error of the regression, which is always smaller than the error of prediction, and defines the average difference between the measured values of the reference standards and the predicted measured values for the standards based on the regression model.

The error of prediction is not a constant value, but is a non-linear function that is dependent on the measured value of a given sample. The error is at its minimum near the mean of the calibration standards' values. The error of prediction becomes slightly larger as the measured value gets larger or smaller than this mean (see Fig. 1).

There is, unfortunately, one major shortcoming of the accepted error of prediction model when applied to the CHARMed PyMca environment. This shortcoming is that the model is strictly applicable only where the calibration data set is homoscedastic. A homoscedastic calibration data set is one in which the deviation of calibration points from the regression line, or variance, is equal across the calibration range. Unfortunately, this assumption is often not valid in the context of XRF calibration; specifically, it is often the case that the absolute variance (not the *relative* variance) is significantly smaller for measurements made near the limits of detection than for measurements of larger values. Figure 2, for example, shows an ordinary least squares (OLS) calibration plot for lead (Fig. 2 (a)) alongside a plot of the associated absolute residuals (Fig. 2 (b)). Here, it is very clear that the absolute variance is highly unequal (heteroscedastic) and is lowest where the certified value is small. In cases such as this, OLS error calculations will tend to overestimate the error in the region where the variance is relatively small, producing confidence intervals that are too wide at low concentrations (Hayes and Cai 2007).

It is clear that some way of constraining the error of prediction at low concentrations, based on the degree of heteroscedasticity in the calibration data set, would lead to a more reliable estimation of error. Unfortunately, modelling the error of prediction in a relatively small data set that exhibits heteroscedasticity is a very complex problem for which no accepted solution appears to exist. A detailed proposal for addressing the conundrum is beyond the scope of this paper, but will be put forward in a forthcoming publication on the subject of error modelling for XRF. In the meantime, it is proposed that the accepted error of prediction calculations described above be taken as the basis for the CHARMed PyMca calibration error model.

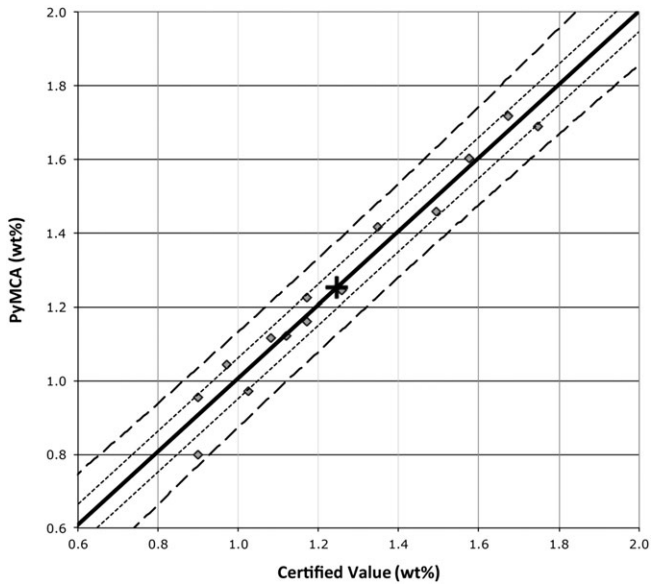

Figure 1 *An example linear calibration regression, showing the calibration points, the standard error of the regression (dotted line), the standard error of prediction at 95% confidence (dashed line) and the mean calibration value* ($\bar{x}$, y_o). *The error of prediction is a non-linear function of the measured value and is at a minimum about the mean value of the calibration standard set.*

Instrumental Reproducibility

Given adequate counting times, most XRF instruments in use in the cultural heritage community are capable of producing consistent results and so errors associated with instrumental reproducibility are generally relatively small compared to the calibration uncertainties. In some instances, however, this error can become significant, usually due to spectral noise associated with low concentrations or short counting times, or due to difficulties associated with peak deconvolution. In the CHARMed PyMca protocol, it is proposed that a model of instrumental error be built independently for each element in the calibration, using the triplicate results for each of the standards as a guide. It is for this reason that the authors recommend collecting the standard spectra on different days and by different operators. Characterization and control of longer-term instrumental error can be addressed by implementing a drift-monitoring program for each instrument.

The most straightforward way to model instrumental reproducibility in this context is to first take the standard deviation of each set of calibrated triplicate measurements for one element, and then calculate the average standard deviation for calibrated results of that element across the calibration range. While the use of only three measurements per sample might be considered less than desirable, when all reference samples are considered, this results in 12 independent estimates of instrumental reproducibility for each element. To estimate the range of variation within which 95% of future repeat measurements would fall, the average standard deviation should be multiplied by 1.96 (the two-tailed z value associated with a 95% confidence interval for a normal distribution). The resulting value can then be considered as a useful estimate of the instrumental reproducibility error and this can, in turn, be propagated with the error of prediction to estimate the complete error budget.

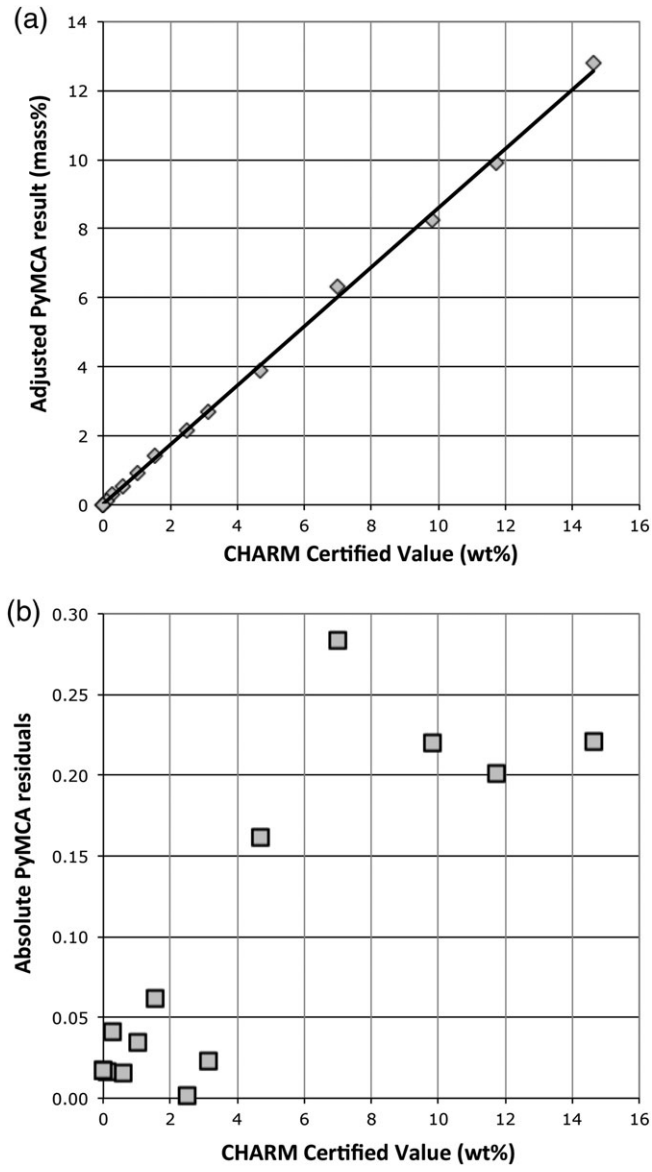

Figure 2 *A typical calibration plot for lead using CHARMed PyMca (using a Bruker Tracer S1 with an Rh tube) (a), and a plot of the associated absolute residuals for the regression (b). The residuals increase with concentration, revealing pronounced heteroscedasticity in the data set.*

Unfortunately, as with the error of prediction, the instrumental error data sets are also frequently heteroscedastic. Specifically, the standard deviation of the 12 triplicate measures for each element often increases as the measured concentration of the element increases. Therefore, a single estimate of instrumental error is often likely to somewhat overestimate error at the low end of the calibration range and to underestimate error in the upper end of the range. This phenomenon is well known and a good discussion is given in Lachance and Claisse (1995, 272–3).

A detailed proposal for constructing a practical model for instrumental reproducibility that accommodates heteroscedasticity is also beyond the scope of this paper, but will also be put forward in a forthcoming publication. In the meantime, it is proposed that the average standard deviation, as described above, be taken as the basis for the CHARMed PyMca instrumental error model.

Application of the Error Models

The error of prediction from the calibration and the instrumental reproducibility error can be considered independent and additive. Therefore, given that they are scaled to the same confidence interval, the accepted method for propagating the two errors is to take the square root of the sum of the squared errors (de Vries and Vrebos 2002). In the vast majority of cases, the instrumental error is significantly smaller than the error of prediction and will contribute less than 5% to the overall error associated with a calibrated result. With short acquisition times and low count rates (e.g., 45 live seconds at 6000 cps), the contribution of instrumental error for some minor elements can rise to as much as 30% of the total.

Formulating a method for applying the propagated error model to the results of new analyses can be accomplished in several ways. A very convenient method that maintains good precision is to simply calculate the propagated error, as described above, for the predicted result of each of the CHARM calibration samples. It is the author's experience that a second-order regression will invariably closely fit a line to these points (with R^2 values of 0.99 and above) and the resulting regression equation can be used to estimate the overall error for any calibrated result in the calibration range. Figure 3 illustrates the overall error model for copper based, on a CHARMed PyMca calibration for an Olympus Delta spectrometer. In this figure, for each predicted (calibrated) result derived from the initial XRF measurements of the copper CHARM set, the difference from the certified value (residual) is plotted as a diamond; the combined error (at 95% confidence) for each result, calculated according to the formulas outlined above, is plotted as a

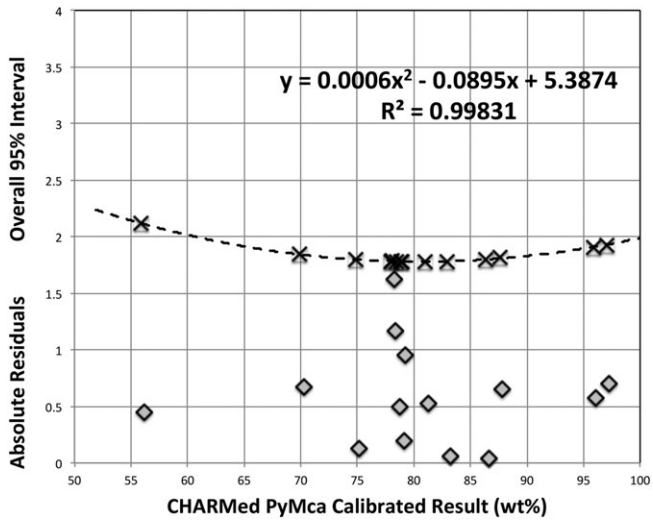

Figure 3 *A plot of the overall error model for copper based on a CHARMed PyMca calibration for an Olympus Delta spectrometer; the absolute residuals are plotted as diamonds; the calculated overall error (at 95% confidence) for each calibration point is plotted as a cross and an excellent approximation of the error for any future result (derived from the quadratic regression of the error points) is shown as the dashed line.*

cross. An excellent approximation of the error for any future result (dashed line) can be made based on the simple quadratic formula derived from the regression of the calculated error points.

Validation

Once a calibration and error model are built for a particular XRF instrument and measurement protocol, it is important to confirm that the model is functioning as designed. The use of a validation set is common for this purpose in many statistical methods. In general, a validation set is a group of well-characterized samples that have not been used in the construction of a statistical or probabilistic model, that are then used to assess the validity of the model. In the context of the CHARMed PyMca procedure, it is recommended that a set of 12 reference materials (certified if possible, but uncertified as necessary) that are not in the copper CHARM set should be selected for this purpose. As much as possible, the validation set should contain the same 15 elements and cover the same range of concentrations as the calibration (CHARM) set, although in practice this is difficult to achieve.

The validation set should, of course, be measured using the same instrumental parameters as the calibration set. The number and temporal distribution of replicates may be adjusted depending on the specific manner in which one wishes to analyse the results. The authors recommend analysing each validation sample three times in immediate succession without moving the sample between analyses. This follows the ASTM protocol for inter-laboratory reproducibility studies, and the results can then be compiled with those from other users to calculate reproducibility statistics such as the method minimum standard deviation (S_M) and the per cent relative reproducibility index (R_{rel}) (ASTM 2003).

The results from the validation set can be plotted directly against the reference values for each sample to confirm that the calibration model is functioning correctly (Fig. 4 (a)). A more useful plot for evaluating the calibration model in detail is a residuals plot overlaid with the error model for the calibration (Fig. 4 (b)). If the calibration model is functioning correctly, the validation set residuals plot will show data points that are roughly symmetrical about the *x*-axis and approximately 95% of the points will lie between the lines that describe the error model. If the data points are weighted above or below the *x*-axis, then an unexplained bias is likely to be present. If the data points are clustered near the *x*-axis and do not approach the error boundaries, then the model is likely to be overestimating the error; in contrast, if significantly more than 95% of the data points fall outside the error model boundaries, then the error model is likely to be underestimating the error.

MECS: A MULTI-ELEMENT CALIBRATION SPREADSHEET

A multi-element calibration spreadsheet (MECS) has been developed in Excel® to facilitate the process of building a functional calibration/error model and implementing the validation procedure. The MECS is written to take advantage of Excel's logical, lookup and statistical functions in order to streamline and largely automate the calibration procedures described above. The MECS also automates the validation procedure so that results from the validation set are displayed and analysed, as recommended above, to confirm that the calibration and error models are valid. In addition, the MECS streamlines the routine quantification of new experimental XRF spectra, automatically applying the CHARM-based calibration and tabulating quantitative results and errors in several customizable report formats as well as generating graphical output of results, also in several formats. On a separate tab, the MECS also presents a calibration report with an

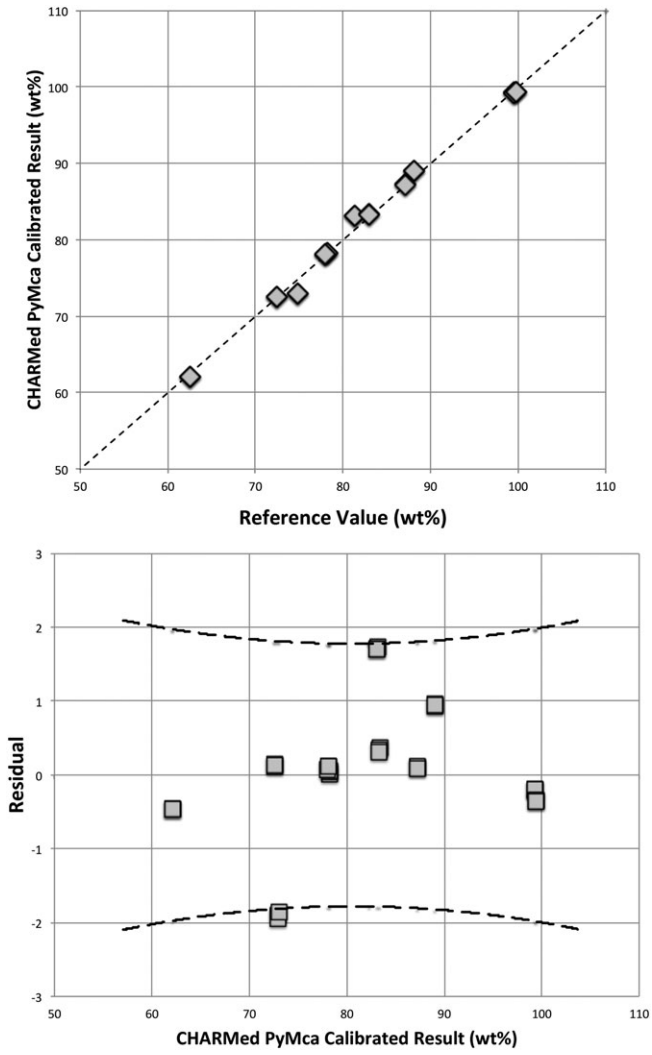

Figure 4 *(a) The calibrated results from the validation set plotted against the certified values for copper (using an Olympus Delta spectrometer with an Rh tube); the dashed line represents a 1:1 correspondence. (b) The validation set residuals plot overlaid with the error model for the calibration (dashed line); if the model is functioning correctly, approximately 95% of the points should lie between the error model lines.*

element-by-element summary table of all the calibration parameters and the constants used for the calculation of all the regression models and the error models that comprise the total calibration scheme.

The MECS described here greatly simplifies and automates calibration and quantification according to the CHARMed PyMca protocol; a complete calibration model can be built from PyMca results in a matter of a few hours and, subsequently, new results can be processed, plotted and tabulated from batched PyMca results in a matter of minutes. The use of an Excel® spreadsheet for this work has its limitations, however, and it is hoped that this MECS might, therefore, be used as a model for the development of new open-source software that could be integrated

with PyMca and used more widely by the art and archaeology community. A copy of the MECS can be downloaded from http://www.getty.edu/museum/conservation/papers/pymca.html.

RESULTS

The CHARMed PyMca protocol was tested by an eight-instrument reproducibility study in which six different instrumental configurations were employed by five analysts. A validation set of 12 reference standards representing a variety of alloy types was analysed using each of the eight instruments and the spectra processed using the protocol. The results of this study were analysed according to the same standard methodology used in the previous 17-instrument reproducibility study (Heginbotham *et al.* 2011). In comparison to the earlier study, the CHARMed PyMca protocol yields significantly improved reproducibility and method sensitivity. On average, for all elements, it resulted in a reduction in the mean % relative reproducibility of 75% (e.g., for lead, the inter-laboratory reproducibility went from ±77% to ±18%). Similarly, using CHARMed PyMca, the method calculated lower limit (the concentration below which relative reproducibility begins to deteriorate rapidly) improved by 51% (e.g., for arsenic, the lower limit dropped from 0.24% to 0.13%).

The CHARMed PyMca protocol also allowed the consistent reporting of 15 elements, in contrast to only eight that were consistently reported in the earlier study. Furthermore, the results were accurate; the interval defined by the group mean result and the reproducibility standard deviation ($S_{R\text{-}95\%}$) contained the standard reference value approximately 92% of the time. The full results of this reproducibility study will be reported in part 2 of this paper.

CONCLUSIONS

CHARMed PyMca certainly will not eliminate all potential problems with ED-XRF analysis of heritage copper alloys. Notably, complications related to surface roughness, gross inhomogeneity, corrosion/patina, surface enhancement and surface depletion will continue to offer challenges to analysts. No doubt, many archaeological copper alloy materials, for example, will continue to be inappropriate candidates for quantitative ED-XRF analysis due to these factors, and analysts will have to remain vigilant about selecting sample sites that are clean, uncorroded, homogeneous and representative of the bulk.

At a minimum, the CHARMed PyMca protocol offers the possibility of dramatically improving reproducibility between laboratories conducting ED-XRF on clean, homogeneous copper alloys. If the full potential of collaborative research on heritage copper alloys is to be fulfilled, researchers must be able to trust in the reproducibility and stated precision of quantitative compositional data generated by collaborating laboratories. ED-XRF faces intrinsic challenges in the analysis of these alloys due to the large number of analytes present, the wide concentration ranges encountered and the high variability of matrix characteristics. The CHARMed PyMca protocol presented here is designed to address these challenges in a manner that is rigorous, readily accessible and fully transparent. It is hoped that by using a shared set of standards, with shared open-access software and a common calibration strategy, this protocol will offer the possibility of sharing and aggregating quantitative data in a manner that is consistent with regard to the elements analysed, well-characterized in terms of precision and demonstrably reproducible.

In the future, it seems reasonable to assume that the essential aspects of this protocol might also be usefully applied to other heritage materials, such as glasses and other metal alloys, if common reference standard sets can be defined. Furthermore, the protocol may also be extended

to other X-ray techniques such as particle-induced X-ray emission (PIXE), or less common variants of ED-XRF such as polycapillary-XRF or synchrotron-μXRF techniques.

ACKNOWLEDGMENTS

The authors are sincerely indebted to all those who have graciously volunteered to trial various iterations of this protocol and offered valuable feedback over the past several years, including Jane Bassett, Federico Caro, Julia Day, Jan Dorscheid, Joseph Godla, Katie Holbrow, Lynn Lee, Arie Pappot and Beth Price. Thanks are also extended to David Bourgarit, Gareth Davies and Aaron Shugar, for their thoughtful commentary.

REFERENCES

ASTM, 2003, *Designation: E 1601—98 (reapproved 2003): standard practice for conducting an interlaboratory study to evaluate the performance of an analytical method*, ASTM International, West Conshohocken, PA.

Barwick, V., 2003, Preparation of calibration curves: a guide to best practice, Report Number LGCVAM2003032, National Measurement Office (UK), Teddington, UK; http://www.nmschembio.org.uk/dm_documents/lgcvam2003032_xsjgl.pdf

Box, G. E. P., and Draper, N. R., 1987, *Empirical model-building and response surfaces*, Wiley, New York.

Bray, P., Cuénod, A., Gosden, C., Hommel, P., Liu, R., and Pollard, A. M., 2016, Form and flow: the 'karmic cycle' of copper, *Journal of Archaeological Science*, **56**, 202–9.

Burgess, C., 2000, *Valid analytical methods and procedures*, The Royal Society of Chemistry, Cambridge.

Caussin, P., 2013, Comparing existing MAC tables—hints to possible developments, *Powder Diffraction*, **28**, 90–4.

Currie, L. A., 1999, Detection and quantification limits: origins and historical overview, *Analytica Chimica Acta*, **391**, 127–34.

de Boer, D. K. G., 1990, Calculation of X-ray fluorescence intensities from bulk and multilayer samples, *X-Ray Spectrometry*, **19**, 145–54.

de Vries, J. L., and Vrebos, B. A. R., 2002, Quantification of infinitely thick specimens by XRF analysis, in *Handbook of X-ray spectrometry* (eds. R. V. Grieken and A. A. Markowicz), 2nd edn, Marcel Dekker, New York.

Ebel, H., 1999, X-ray tube spectra, *X-Ray Spectrometry*, **28**, 255–66.

Fernandez, J. E., Scot, V., Verardi, L., and Salvat, F., 2014, Detailed calculation of inner-shell impact ionization to use in photon transport codes, *Radiation Physics and Chemistry*, **95**, 22–5.

Fernandez, J. E., Scot, V., Verardi, L., and Salvat, F., 2013, Electron contribution to photon transport in coupled photon–electron problems: inner-shell impact ionization correction to XRF, *X-Ray Spectrometry*, **42**, 189–96.

Frahm, E., 2013, Is obsidian sourcing about geochemistry or archaeology? A reply to Speakman and Shackley, *Journal of Archaeological Science*, **40**, 1444–8.

Frank, C., and Pernicka, E., 2012, Copper artefacts of the Mondsee group and their possible sources, in *Lake dwellings after Robert Munro: proceedings from the Munro International Seminar: the lake dwellings of Europe, 22nd and 23rd October 2010, University of Edinburgh* (eds. M. S. Midgley and J. Sanders), Sidestone Press, Leiden, Leiden.

Goodale, N., Bailey, D. G., Jones, G. T., Prescott, C., Scholz, E., Stagliano, N., and Lewis, C., 2012, pXRF: a study of inter-instrument performance, *Journal of Archaeological Science*, **39**, 875–83.

Hayes, A., and Cai, L., 2007, Using heteroskedasticity-consistent standard error estimators in OLS regression: an introduction and software implementation, *Behavior Research Methods*, **39**, 709–22.

Heginbotham, A., Bassett, J., Bourgarit, D., Eveleigh, C., Glinsman, L., Hook, D., Smith, D., Speakman, R. J., Shugar, A., and Van Langh, R., 2015, The copper CHARM set: a new set of certified reference materials for the standardization of quantitative X-ray fluorescence analysis of heritage copper alloys, *Archaeometry*, **57**, 856–68.

Heginbotham, A., Bezur, A., Bouchard, M., Davis, J. M., Eremin, K., Frantz, J. H., Glinsman, L., Hayek, L.-A., Hook, D., Kantarelou, V., Karydas, A. G., Lee, L., Mass, J., Matsen, C., Mccarthy, B., Mcgath, M., Shugar, A., Sirois, J., Smith, D., and Speakman, R. J., 2011, An evaluation of inter-laboratory reproducibility for quantitative XRF of historic copper alloys, in *Metal 2010: proceedings of the interim meeting of the ICOM–CC Metal Working Group, October 11–15, 2010, Charleston, South Carolina, USA* (eds. P. Mardikian, C. Chemello, C. Watters, and P. Hull), Clemson University Press, Clemson, SC.

Institute for Reference Materials and Measurements, 2010, *Frequently asked questions on calibration*, Joint Research Center, European Commission, Brussels.

Lachance, G. R., and Claisse, F., 1995, *Quantitative X-ray fluorescence analysis: theory and application*, Wiley, Chichester.

Mantler, M., Willis, J. P., Lachance, G. R., Vrebos, B. A. R., Mauser, K.-E., Kawahara, N., Rousseau, R. M., and Brouwer, P. N., 2006, Quantitative analysis, in *Handbook of practical X-ray fluorescence analysis* (eds. B. Beckhoff, B. Kanngiesse, N. Langhoff, R. Wedell, and H. Wolff), Springer-Verlag, Berlin.

Rehren, T., and Freestone, I. C., 2015, Ancient glass: from kaleidoscope to crystal ball, *Journal of Archaeological Science*, **56**, 233–41.

Rowe, H., Hughes, N., and Robinson, K., 2012, The quantification and application of handheld energy-dispersive X-ray fluorescence (ED-XRF) in mudrock chemostratigraphy and geochemistry, *Chemical Geology*, **324–5**, 122–31.

Schoonjans, T., Solé, V. A., Vincze, L., Sanchez del Rio, M., Appel, K., and Ferrero, C., 2013, A general Monte Carlo simulation of energy-dispersive X-ray fluorescence spectrometers—Part 6. Quantification through iterative simulations, *Spectrochimica Acta, Part B: Atomic Spectroscopy*, **82**, 36–41.

Solé, V. A., Papillon, E., Cotte, M., Walter, P., and Susini, J., 2007, A multiplatform code for the analysis of energy-dispersive X-ray fluorescence spectra, *Spectrochimica Acta, Part B*, **62**, 63–8.

Solé, V. A., Schoonjans, T., Karydas, A. G., Guijarro, M., Vincze, L., and Heginbotham, A. (in preparation), The fisx software library for analytical X-ray fluorescence calculations. Source code repository available at https://github.com/vasole/fisx

Speakman, R. J., and Shackley, M. S., 2013, Silo science and portable XRF in archaeology: a response to Frahm, *Journal of Archaeological Science*, **40**, 1435–43.

Archaeometry published by John Wiley & Sons Ltd on behalf of University of Oxford, *Archaeometry* **59**, 4 (2017) 714–730