From Papertape Input to 'Forensic Crystallography': A Short History of the Program PLATON

from the Kenneth N. Trueblood Lecture by Ton Spek

A.L. Spek is an Emeritus Professor at the Bijvoet Centre for Biomolecular Research, Utrecht University, The Netherlands

I am 24 years younger than Kenneth N. Trueblood (1920-1998) whose name is associated with the prestigious Trueblood award that I proudly received in Chicago. Notwithstanding a one generation age difference, I share with Ken doing software development and having done that in the early and adventurous days on, from the current perspective, rather primitive computing facilities. I have been fortunate to have met him and other early software developers at various meetings. In the following I will sketch my more than 40 years journey in small molecule crystallography that culminated in the development of the program PLATON that implements much of what I have learned on the way including the Shomaker & Trueblood TLS rigid-body model. PLATON in its structure validation incarnation turned out to be instrumental in the recent uncovering of a massive and saddening fraud case with papers published in *Acta Cryst*. Section E. Nobody was expecting such a thing to happen on such a large scale in the non-competitive chemical crystallography world.

My scientific CV is simple: I studied, obtained my PhD and worked until my official retirement in November 2009 at the same university. Back in 1966 I started as a student in crystallography in the Laboratory for Crystal and Structural Chemistry at Utrecht University that was at that time headed by A.F. Peerdeman. Peerdeman was the successor of J.M.Bijvoet who had retired in 1962. He was co-author of the famous Bijvoet, Peerdeman & van Bommel *Nature* paper on absolute configuration determination. After WWII, Bijvoet had started a new laboratory in a stately house (used by the Gestapo during WWII) close to the center of the city of Utrecht. Part of the house was his private domain. After his retirement, he still kept a pied-a-terre in the former private quarters for when he was in Utrecht to look-up literature in the library for a book he was writing. As a student, I shared the family bedroom - in its double function as student room. We were expected to find a place elsewhere to work when Bijvoet was in town. The laboratory moved in 1973 to a new building in the university campus outside the city.

The Crystal Palace, as the laboratory was often referred to, was also the home of the first generations of computing platforms within the University of Utrecht (Zebra and Electrologica X1 respectively). In 1966 computing had moved to a new university computing center elsewhere in the city. Computing was from then on until the early 1970's done on an Algol language specific X8 computer from the Dutch company Electrologica which isolated us from the FORTRAN world elsewhere. Nearly all crystallographic software had to be developed in-house. FORTRAN programs such as ORTEP and later on MULTAN could not be used. You had to be a software developer as well as a crystallographer in those days. Processing on the X8 was essentially a one job at a time operation with all input via paper tape and output over a line printer or Calcomp plotter. Computing jobs were run by an operator during daytime shifts. Most of our serious crystallographic work had to be done during the once-a-week 13 hour nightshift when we as crystallographers had the university computer for ourselves. Half of the staff and students stayed overnight to run their own jobs in turn. We scientists were all at that time also programmers and computer operators. A block-diagonal least squares cycle might take one hour and an E-map ten minutes. The preparation of programs and program input was done on a so-called Flexowriter. This very noisy electrical typewriter was also often used as output medium. Editing was done with a pair of scissors to cut out unwanted material from the source code or data and adhesive tape to glue a substitute into the paper tape. A lot of time was spent on optimizing memory and execution time requirements. Program and data had to fit in sixteen thousand machine words. One of my early achievements was the creation of an alternative algorithm as a practical substitute for an elegant piece of code by a professional programmer, bringing execution time down from many hours to one minute.

My supervisor, Jan A. Kanters, gave me what turned out to be an interesting assignment to work on. He handed me a batch of white crystals with unknown composition, code named M200. The assignment was to find out what the composition and structure might be, using single crystal x-ray techniques only. In hindsight his assignment very much determined the rest of my career. Preliminary investigations done with film data suggested the centro-symmetric space group P1bar. A 2D Patterson synthesis based on laboriously collected integrated zero layer Weissenberg intensity data subsequently suggested a light atom structure. This blocked any further analysis with the 2D data. Eventually a three-dimensional data set was collected with an Enraf-Nonius AD3 diffractometer that was operated via an instruction patch panel and setting angles for reflections on paper tape. The latter tape had to be created on the X8 on the basis of an orientation matrix. This was a two week data collection for an eleven non-hydrogen atom compound! It took me half a year to finally solve the structure with the 3D data assuming equal scattering type atoms. The laboratory had a tradition in Direct Methods (Paul Beurskens, Ad de Vries, Jan Kroon, Henk Krabbendam). However, all available software failed to solve my structure. These were pre-MULTAN days! In the end I had to write my own Direct Methods program that solved the triclinic structure and subsequently many other unsolved structures that were hanging around in the lab. AUDICE. as the program was named, was one of several Symbolic Addition Method programs that were developed in that period. Its specialty was that at the start of the evaluation of the triple product phase relations with strong indications for a positive sign, 27 symbols were assigned to strongly interacting starting reflections rather than just three as was the case in many other approaches. The number 27 is not arbitrary but represents the number of bits of X8 computer words. Eventually, by eliminating 24 symbols based on multiple symbolic 'indications', 8 solutions were produced with figures of merit. The structure analysis showed that the triple bond in the original dicarboxylic acid had reacted with the methanol solvent of crystallization. The crystal structure of M200 was published subsequently. Unfortunately attempts to publish the algorithms that were used in the program AUDICE in *Acta Cryst*. were blocked by a referee requirement that performance be compared with performance on non-ALGOL (so-called real ..) platforms. That was a killer at that time. Anyway AUDICE was superseded by the program MULTAN (FORTRAN) when the University eventually moved to a multi-user Control Data FORTRAN standard mainframe in the early 1970s. The complete structure determination process that took over half a year has now been automated. M200 is solved and refined in a matter of seconds on current hardware such as my MacBook Pro with the SYSTEM-S tool in PLATON when run in the so-called *No-Questions-Asked* mode.

Multiple meetings and schools were organized in the 70's with Direct Methods (software and theory) as the major subject. Examples are the NATO schools in Parma, Italy (above) and York (UK), the schools in Erice, Italy in 1978 (see at right), and the meetings at the Medical Foundation (Buffalo) and Gottingen (Germany). Important and inspiring were the CECAM workshops on Direct Methods in the early 70's in Orsay (near Paris) around a big European IBM-360 with lectures by Herbert Hauptman (5 weeks (!) that brought together people working on current issues related to Direct Methods). Among the participants were Gabriel Germain, Peter Main, Ricardo Destro, Davide Viterbo and Henk Krabbendam. The program MULTAN was finalized there including interfaces to high end interactive graphics. Coming from the limited X8 & paper tape world into the multi-processing, FORTRAN and the punched card world was a culture shock.

In 1971, a national single crystal service facility was set up in Utrecht, with me to make it all happen. I kept that position for 38 years until my emeritus status in 2009. The project is now continued by my former co-worker Martin Lutz. My last postdoc was Maxime Siegler, now staff crystallographer at John Hopkins University. The program PLATON is a side product of that national facility. A lot of free time went into it as a job related hobby. Its development has never been funded explicitly! Work on PLATON started in 1980 in order to manage the analysis of the growing number of structure determination projects. It was to replace an earlier ALGOL suite of programs and was designed to interface with SHELX76. The idea was to produce with a single 'CALC ALL' instruction an exhaustive listing of all relevant derived geometry, including ring puckering analysis, least-squares planes etc., to be handed over to our clients as a structure report. Over time numerous additional tools have been added on the basis of our service needs, our local research and valuable ideas from external users. PLATON has become, in combination with SHELX(L/S), DIRDIF and SIR one of the major working horses of our national service.

PLATON is purposely designed as a single program, as independent as possible from external libraries. The tools available in PLATON are shown as clickable options on the opening window of the program. Examples are ADDSYM for detection of missed symmetry, TwinRotMat for automatic twinning detection, SQUEEZE for handling disordered solvents, SYSTEM S for guided/automated structure determination, FLIPPER as a new approach for solving the phase problem and CHECKCIF for structure validation.

Reporting structures in the correct space group is a major issue. Dick Marsh has reported numerous cases where a description in a higher symmetry space group was in order. Yvon LePage published an excellent algorithm (MISSYM) for detecting possible higher symmetry elements in a structure. The actual implementation of that additional symmetry is left to the analyst. ADDSYM also implements that step and provides the proper space group name and associated transformations. In that way, the complete CSD can be examined automatically for possible missed symmetry cases. In response to one of Dick's

space group error papers, I wrote to him to ask whether he was interested in my list of structures needing detailed inspection. He was indeed and was amused to find out that one of his previously published corrections was 'Marshed' again in that a still higher space group symmetry was found.

The SQUEEZE tool was created to make the publication of the structure of a pharmaceutical that was already hard to crystallize possible. The structure exhibited infinite channels filled with disordered solvent. The tool consists of two parts. In the first part the solvent accessible volume in a structure is identified. In the second part that volume is used as a mask on the electron density found in that region. Iterative back-Fourier transformation of that density provides the solvent contribution to the calculated structure factors.

A busy author or referee can easily miss problems with a structure. Increasingly black-box style analyses done by non-experts are being published. The number of referees and experts available for detailed examination of the exploding number of structure reports is quite limited. It is easy to hide problems from the experts with a ball-and-stick style illustration. Sadly, fraudulant results and structures have been identified in the literature that contaminate the assumed solid information archived in the CSD. Automated Structure Validation as a solution for this problem was pioneered and 'pushed' by Syd Hall as editor of *Acta Cryst*. Section C with the creation of the CIF standard for data archival and exchange (Hall et al., (1991) *Acta Cryst*. A47, 655-685. He also encouraged George Sheldrick to adopt CIF for the then new SHELX97 refinement program. Subsequently he made CIF the *Acta Cryst*. Section C submission standard and set up early CIF checking procedures for submitted CIF's. I was invited to include PLATON checking tools such as ADDSYM and VOID search for missed solvent accessible voids. Over time several hundreds of new ALERTS were introduced on the basis of issues I detected as *Acta Cryst*. Section C Co-Editor. Validation was made into a standard WEB-based tool by the IUCr Chester staff and strongly imposed by the next Section Editors George Ferguson and Tony Linden. The validation scheme has been very successful for *Acta Cryst*. Sections C & E in setting standards for quality and reliability. The missed symmetry problem has effectively been solved for the IUCr journals though unfortunately not yet for other journals. There are still numerous 'Marshable' structures published as Dick Marsh keeps showing although most major chemical journals now have some form of a validation scheme implemented.

The IUCr has recently made one step further with FCF-validation. *Acta Cryst*. is unique in requiring that reflection data have to be archived for published papers in computer readable format. This is standard in the bio-crystallography world but surprisingly not in chemical crystallography. When validation of the structure factor data is included then sloppy and even fraudulent practices become obvious. Errors are easily made and unfortunately not always discernable from fraud in the absence of deposited reflection data. It took some time to discover a pattern of systematic fraud. Wrong element type assignments can be caused as part of an incorrect analysis of an unintended reaction product. Alternative element types can also be (and have been) substituted deliberately in order to create 'new publishable' structures. Reported and calculated R-values differing in the first relevant digit!? have been detected, obviously meant to 'clean up' the validation list of ALERTS. Until recently, nobody seems to have looked seriously at the other structures published by the authors of a strangely incorrect structure. Doing so and as part of the testing of FCF validation software, a large fraud was detected with papers published in *Acta Cryst*. Section E around 2007. Over 100 structures have now been retracted and marked as such in the CSD. A whole series of 'isomorphous' (often chemically impossible) structures was detected for an already published (correct) structure. The data sets of different structures could be shown to be identical. Similar series have now been detected for coordination complexes (transition metals and lanthanides). ***How could referees have let those pass?***

Recently, it was realized that there is an 'age-concern' issue in that many software developers are retiring with only a limited next generation. IUCr Computing Schools organized in Siena in 2005 (see below), and in Kyoto (2008) addressed this issue. Lachlan Cranswick was the major force in that project.

From left: Ton Spek, Tony Linden (current Section Editor of Acta Cryst. C), George Ferguson (previous Section Editor), and Aggie Spek in Zurich.

Finally I would like to thank the selection commission for this award; my former co-workers in the project over 38 years and in particular my successor Martin Lutz; Louis Farrugia for following my frequent updates of native UNIX PLATON with his MS-Windows implementation; the users of the software for communicating new ideas and bug reports; Lachlan Cranswick, who sadly is no longer with us, for promoting my software, and my wife for enduring me with my notebook.

Ton Spek