Surf. Sci. **604**, 1544–1547 (2010) http://dx.doi.org/10.1016/j.susc.2010.06.017

Surface Science Prospective

A standard format for reporting atomic positions: further needs and options

M.A. Van Hove[1], K. Hermann[2], P.R. Watson[3], D.P. Woodruff[4], S.Y. Tong[5], R.D. Diehl[6], K. Heinz[7], C. Minot[8], H. Tochihara[9]

[1] Department of Physics and Materials Science, City University of Hong Kong, Hong Kong, China
[2] Theory Department, Fritz-Haber-Institut der MPG, 14195 Berlin, Germany
[3] Department of Chemistry, Oregon State University, Corvallis, OR 97330, USA
[4] Department of Physics, University of Warwick, Coventry CV4 7AL, UK
[5] Department of Physics, University of Hong Kong, Pokfulam, Hong Kong, China
[6] Department of Physics, Penn State University, University Park, PA 16802, USA
[7] Department of Physics, University of Erlangen-Nürnberg, Staudtstraße 7, 91058 Erlangen, Germany
[8] Laboratoire de Chimie Théorique, Université Pierre & Marie Curie, Paris 6, CNRS, UMR7616, Case 137, 4 place Jussieu, Paris, F-75252 Cedex France
[9] Department of Molecular and Material Science, Kyushu University, Kasuga, Fukuoka 816-8580, Japan

Abstract

In response to a recent Surface Science Editorial [1] and accompanying Prospective article [2], we propose that the community of surface scientists and nanoscientists select a common data format for reporting experimental data and atomic coordinates, and have it be accepted by relevant journals.

1. Introduction

A recent *Surface Science* Editorial [1] emphasized that "whenever a crystal structure is being discussed, our editorial policy requires that all the atomic positions be made accessible to the reader". It added: "Any experimental data used to determine that crystal structure should also be made available as Supplementary Material." We wholeheartedly agree with this general request.

A Prospective article [2] accompanying that Editorial proposed "that the Surface Science community adopt the same standard format for reporting these as is already widely used in bulk crystallography publications, namely the inclusion of a Crystallographic Information Format file (or CIF file) as supporting information." We wish to clarify that many data formats are available for archiving structural as well as other physical and chemical information, although at present none of these, including CIF, is general enough for the current needs.

For example, as stated in the mentioned Prospective article, CIF needs extensions to cover a wider range of techniques used in surface science. More importantly, surface science has evolved into nanoscience, which includes many lower-dimensional structures worth reporting: unfortunately, no single current data format is flexible enough to handle atomic structures in 3 dimensions (e.g. periodic bulk crystals), 2 dimensions (e.g. periodic free surfaces, interfaces), 1 dimension (e.g. nanowires, nanotubes, line defects), as well as 0 dimension (e.g. nanoparticles), on a common basis. Other interesting systems that don't fit existing formats are incommensurate structures (e.g. graphene on a crystalline substrate and interfaces between two semi-infinite materials) and disordered (non-periodic) structures. Many current formats do allow some limited flexibility through an artificial 3-dimensional supercell. Here the structure of the supercell as a whole is used to characterize the complete system, often in an approximate way, for example using a repeated-slab geometry. However, this may mask important details of a structure (e.g. the semi-infinite 3D substrate of a 2D surface, or multiple layer-dependent 2D unit cells, or the isolation of an interface). Also, no data format to our knowledge is set up to report multiple techniques used in a single structure determination, e.g. low-energy electron diffraction and density functional theory used together to narrow down the range of possible structures.

We therefore wish in this Prospective article to discuss in more detail the needs that we perceive in selecting a suitable data format to achieve the stated goal and some options that we have for moving forward.

2. Needs

We start by listing what would ideally be needed to conveniently make available experimental and theoretical results, including structural data such as atomic positions:

1) Relevant journal and book publishers should enforce the same policy and the same format.

2) The ideal data format would allow specifying the methods used in enough detail to enable reproducing the results.

3) The ideal data format would allow reporting the use of multiple techniques in a single structure determination (e.g. determination from experiment and from first principles).

4) The ideal data format would allow presenting coordinates for all 3-dimensional and lower-dimensional structures, including combinations thereof (e.g. a carbon nanotube or a graphene sheet adsorbed incommensurately on a single-crystal surface).

5) The ideal data format should allow incomplete structural results, as many studies provide very valuable yet partial data (e.g. H atom positions may not be determined, or an internal molecular structure may be determined while its adsorption site on a surface may not be).

6) All data would ideally be set in a unique machine-readable format, preferably using human-readable tags. Corresponding formats should be described in the open literature and should be free of copyright protection issues.

7) All past published results would be converted to the selected format.

A few comments on these requirements are appropriate:

1) To achieve enforcement by relevant publishers would require a priori consensus on the appropriate data format within the community of experimental and theoretical crystallographers, surface scientists and nanoscientists.

2) There is an issue of amount of detail needed when publishing the methodology of structure determination: should a calculation, for example, be reproducible by the reader with 6-digit accuracy (requiring much computational detail) or only be roughly described (by only specifying the methods used with corresponding input parameters)?

5) Incomplete structure determinations are in fact rather common in surface crystallography, and occur in the Surface Structure Database. Often, parts of the structure are simply assumed, e.g. the 2D lattice constants of a surface and the bulk structure below a surface. Some atom positions may not be found (hydrogen is typical), or the adsorption site of a molecule may be undetermined even though the internal molecular structure is found. The Prospective article [2] includes an example in which the bulk structure was not included because that structure refinement only used superstructure spots. In nanostructures, one must expect many more partial structure determinations. In terms of formats, missing atoms are easily dealt with by simple omission (as is done in SSD and in the CIF example in Ref. [2]). Assumed structural parts can be included in the data file but should be identified as being assumed or fixed and "not determined". Undetermined quantities like adsorption sites can be handled by assuming a value, e.g. a specific site, which assumption should be explicitly stated.

6) While the Surface Science editorial policy currently accepts the listing of atomic coordinates in tables in the body of an article (if not too voluminous), for easy accessibility and for uniformity it is desirable that a machine-readable dataset be provided. Journals should also be prepared to handle voluminous experimental datasets, especially in the form of 2-dimensional numerical diffraction patterns, scanning tunneling microscope images, 3-dimensional holographic datasets, sets of graphs, etc.

7) Producing data files for past surface structure determinations is clearly impractical, especially as most scientists involved in their initial creation are no longer active in the field. However, surface structures included in the Surface Structure Database [3] are accessible for conversion, but would require a conversion utility.

Generally, these requirements speak for a single universal format. This would also minimize inter-conversions of data between formats that take time and risk errors. This could be done by suitably generalizing existing bulk structure formats, such as the CIF file format proposed in the recent Prospective article [2]. However, a careful implementation would have to consider and allow many additional features introduced by novel nanosystems. This may require a redesign of the format beyond simple patches to the existing version.

Another option is to follow the format used in the Surface Structure Database (SSD) [3, 4], which was documented in the open literature [5] and on the internet [6], since it was designed specifically for surface structures. Unfortunately, although it was available when most surface structures were being solved, this format has not taken root, and was not even mentioned in the Prospective article [2] addressing this issue. In fact, updating the database itself was discontinued due to the fall-off in published surface structures. While some past authors continue to publish structures, new authors have appeared in the field of surface structure. However, of the 30 most prolific contributing authors to SSD up to 2004, only 6 are now still publishing structures to our knowledge.. It is therefore not clear why a different format, which is not yet ready for surfaces and less specific to surfaces, would fare better now.

It seems that a data format will only be used by scientists if publishers compel them to do so. It also appears that an update in formats is needed to address the types of structure which are studied today and in the foreseeable future: these include prominently nanostructures and complex surfaces of all sorts, including molecular networks, i.e. structures that very often have lower dimensionality and order than two-dimensionally periodic surfaces.

3. Format options

We here discuss and compare some existing formats and options for the future.

The first requirement for publishing experimental and structural data in machine-readable form is a data format, preferably universal. Practically speaking, one must probably allow a simple free format without systematic rules for experimental data and for (x,y,z) coordinates; this would make submission of data very easy, especially for simpler structures and for those scientists who do not want to deal with format issues. The free format, by its ad-hoc nature, is very flexible and can represent essentially any type of structure. However, due to its lack of a strict format definition, it becomes case-dependent and makes inter-conversion of data difficult and error-prone, if not impossible.

For experimental data, it is likely that each experimental technique (e.g. x-ray diffraction, low-energy electron diffraction, photoelectron diffraction, electron holography) will require its own format or formats, perhaps due to the particular software used to collect the data. These formats should be described in the open literature and be free of copyright protection issues. Ideally, data of these formats should be character based, i.e. accessible with standard text editors.

For atomic coordinates, many formats, both of commercial origin or from open sources, are already available. Examples are CIF, XYZ, PDB, INS, RES, ICSD, FDAT, MDL, CSSR, SSD and SURVIS. Most of these are designed for molecular or 3D periodic structures, or, in the case of SSD, for 2D and semi-infinite 3D periodic structures. The SURVIS format [6] allows 0D, 1D, 2D, and 3D periodic structures, however, only in numerical form without textual comments. We assume that most other formats are equivalent to either the free or CIF formats.

Clearly, from a conceptual point of view, any of the above mentioned data formats can be used to describe a general system with 0- to 3-dimensional components, if the supercell approach is applied. However, this strategy may hide the different components in a structure file, thus, making a physical description less obvious.

In the following we therefore restrict ourselves to a comparison of the capabilities of the CIF, SSD and SURVIS formats. Table 1 compares these formats for a few types of relevant structures.

It becomes rapidly apparent from Table 1 that none of these formats qualifies as the "universal format". One could design a single new universal data format to replace and generalize all the others. Obviously, this would take considerable consultation and development. In addition, past experience has shown that scientists will always prefer to use formats with which they are familiar and use new formats only if they see the need for it.

A simpler alternative is to allow a "hybrid universal format", in which different formats may coexist in the same structure file for different parts of the structure. As an example, for a 2D periodic surface with a 1D periodic nanotube adsorbed on it, the 2D surface could be described in one format and the nanotube in another format.

This "hybrid universal format" could, conceptually, take the following form, consisting, in a full version, of 6 parts (parts 3 - 6 are optional and depend on the system):

1) General description: verbal characterization of structure, theoretical, experimental procedures, dates, collaborators, publications, links to other publications, links to data sets, legal comments.

2) Format descriptor for each structural component: the overall structure consists of
   - n_0 different 0D cluster components: possible formats XYZ, PDB, INS, RES, ICSD, FDAT, MDL, CSSR, etc.
   - n_1 different 1D chain components: possible formats SSD, CIF, PDB, SURVIS, etc.
   - n_2 different 2D layer components: possible formats SSD, CIF, PDB, SURVIS, etc.
   - n_3 different 3D crystal components: possible formats SSD, CIF, PDB, SURVIS, etc.

3) 0D component(s) (molecules, clusters, nanoparticles) with, for each component (in its own format):
   - origin offset
   - common/individual symmetry, coordinates (x, y, z, element)
   - additional comments

4) 1D component(s) (cluster chains, nanowires) with, for each component (in its own format):
   - origin offset
   - spatial confinement
   - 1D lattice vector R_1
   - common/individual symmetry, structure: coordinates (x, y, z, element) of 1D unit cell
   - additional comments

5) 2D component(s) (cluster layers, monolayers) with, for each component (in its own format):
   - origin offset
   - spatial confinement
   - 2D lattice vectors R_1, R_2
   - common/individual symmetry, structure: coordinates (x, y, z, element) of 2D unit cell
   - additional comments

6) 3D component(s) (3D crystal components) with, for each component (in its own format):
   - origin offset
   - spatial confinement
   - 3D lattice vectors R_1, R_2, R_3
   - common/individual symmetry, structure: coordinates (x, y, z, element) of 3D unit cell
   - additional comments

The hybrid universal format would be most general and at the same time has the virtue that, in simple cases, scientists could use their own choice of an existing format without further modifications except for the format definition at the beginning of a data collection.

4. Path forward

It appears to us that achieving a systematic inclusion of experimental data and atomic coordinates requires undertaking the three following efforts:

1) The scientific community needs to agree on a single universal format or a hybrid universal format (the latter as proposed above). We consider the hybrid format to be conceptually simpler, faster to implement, and able to reach acceptance more easily.

2) Volunteers need to step forward to design, implement and maintain the chosen format.

3) Relevant journals must be approached and convinced to require submission of such data in the same chosen format.

These are challenging objectives, but ones that would be hugely advantageous to future progress in structural studies. Of course, in the meantime *any* improvement in making accessible more quantitative information on structural studies would be of benefit.

To address steps 1) and 2), we invite opinions on this proposal or suggestions to be submitted to the Editor-in-Chief of *Surface Science* (Prof. Charles Campbell, at surfacesci@chem.washington.edu), with his encouragement.

References

[1] Editorial, Surface Science 604 (2010) 877.

[2] L.D. Marks, Surface Science 604 (2010) 878.

[3] Surface Structure Database (SSD) Version 5, P.R. Watson, M.A. Van Hove and K. Hermann, NIST Surface Structure Database Ver. 5.0, NIST Standard Reference Data Program, Gaithersburg, MD, USA (2004). http://www.nist.gov/srd/nist42.htm

[4] M. A. Van Hove, K. Hermann und P. R. Watson, Acta Cryst. B 58 (2002) 338.

[5] P. R. Watson, M. A. Van Hove, and K. Hermann, "Atlas of Surface Crystallography based on the NIST Surface Structure Database (SSD)", J. of Phys. Chem. Ref. Data, Monograph No. 5, ACS 1994.

[6] SSDIN software package for submitting structure data to the Surface Structure Database (SSD) including the surface visualizer SURVIS. The SSD and SURVIS file formats are described in the interactive help files of the SSD and SURVIS software as well as on the SSDIN web page at http://www.fhi-berlin.mpg.de/KHsoftware/ssdin5/index.html from where the package can be downloaded.

[7] S.R. Hall, F.H. Allen, I.D. Brown, Acta Crystallographica Section A: Foundations 47 (6) (1991) 655.

[8] S. Hall, Acta Crystallographica Section A: Foundations 54 (6 Part 1) (1998) 820.

[9] I.D. Brown, B. McMahon, Acta Crystallographica Section B: Structural Science 58 (3 Part 1) (2002) 317.

[10] International Tables for Crystallography Volume G: Definition and Exchange of Crystallographic Data, IUCR, Chester, 2005 Vol. G.

Table 1. Capabilities of different data formats (CIF, SSD, SURVIS) for various relevant structure types.

Data Format vs. Structure Type	**CIF**	**SSD**	**SURVIS**
	[7-10]	[3,4,6]	[6]
0D and 1D nanostructures (periodic or finite-sized)	Yes with supercells, but needs generalizations for more suitable representations	Yes with supercells, but needs generalizations for more suitable representations	Yes with or without supercells
2D surfaces (periodic or not)	Yes with supercells, but needs generalizations for more suitable representations (e.g. semi-infinite bulk, incommensurate or disordered layers)	Yes, with limited capability for non-periodic structures (incommensurate superlattices, lattice-gas disorder allowed)	Yes with or without supercells, lattice-gas disorder allowed
Combination of 2D surface and 1D or 0D adsorbate	Only if surface and adsorbate fit in a common supercell, and if ignoring the semi-infinite bulk	Yes, with limitations as for 2D surfaces	Yes
Multiple 2D unit cells (e.g. (1x1) and superlattice)	Yes but without indication which atoms satisfy (1x1) periodicity	Yes	Yes
3D nanostructures (periodic or finite-sized)	Yes, with supercell if finite-sized	No	Yes
Theory details	Yes	Yes, but limited	No
Experimental data	Yes, but large datasets may challenge journals	No	No
Use of human-readable data names	Yes	No	No
Conversion routines	Some exist	A few exist	A few exist