Representation of multiple and aromatic bonds by a π-electron formalism, which enables an efficient and elegant handling, SDF list no. 1[14]
Support of stereochemistry
Characterization of the stereocenters by parity vectors
Tautomers are registered separately
Cross referencing of stereoisomers
Cross referencing of tautomeric structures (not implemented yet)
Routines for consistency checking

ACKNOWLEDGMENT

I thank my colleagues who have participated in the design and development of this system, especially S. Welford, T. Cieplak, M. Heinen, and B. Roth (Chemplex GmbH) and all the Beilstein chemists who worked out the chemical concepts and offered continuous help and control. The project was supported by the German Ministry of Research and Technology.

REFERENCES AND NOTES

(1) Dittmar, P. G.; Stobaugh, R. E.; Watson, C. E. The Chemical Abstracts Service Chemical Registry System. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 111-121.
(2) Ryan, A. W.; Stobaugh, R. E. The Chemical Abstracts Service Chemical Registry System. 9. Input Structure Conventions. *J. Chem. Inf. Comput. Sci.* **1982**, *22*, 22-28.
(3) Domokos, L.; Jochum, C.; Wittig, G. Data in Beilstein-Online. *Mikrochim. Acta* **1986**, *2*, 423-429.
(4) Data Structure of the Beilstein Database, internal documentation, available from the Beilstein Institute.
(5) The host independent memory resident chemical structure query editor MOLKICK. *Beilstein Brief*, **1988**, 2.
(6) Domokos, L.; Goebels, L. *Der Computer als Nomenklatur-Struktur-Dolmetscher, Tagungsbericht, GDCH 3.* Vortragstagung: Würzburg, 1986.
(7) Rohbeck, H. G. Representation of Structure Description Arranged Linearly. In *Software Developments in Chemistry 5*; Gmehling, J., Ed.; Springer Verlag: New York (in press).
(8) Welford, S. M. Structure Registration for Beilstein Online, Second International Meeting on Chemical Structures, Nordwijkhout, 1990; Warr, W. A., Ed.; Springer Verlag: Berlin (in press).
(9) Welford, S. M. Tautomer Processing in the Beilstein Registry System. In *Software Entwicklung in der Chemie 2*; Gasteiger, J., Ed.; Springer Verlag: New York, 1988; pp 35-43.
(10) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures. *J. Chem. Doc.* **1965**, *5*, 107-113.
(11) SDF and BRCT, internal documentation, available from the Beilstein Institute.
(12) Jochum, C. Building Structure-Oriented Numerical Factual Databases: The Beilstein Example. *World Patent Information* **1987**, *9*, 147-151.
(13) Hicks, M. G.; Jochum, C. Performance Comparison of the MACCS, DARC, HTSS, CAS Registry MVSSS, and S4 Substructure Search Systems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 191-199.
(14) Gasteiger, J. A representation of π-systems for efficient computer manipulation. *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 111-115.

The STAR File: A New Format for Electronic Data Transfer and Archiving

SYDNEY R. HALL
Crystallography Centre, University of Western Australia, Nedlands 6009, Australia

Received October 2, 1990

A new type of format is proposed for the computer archiving and electronic transmission of text and numerical data. The Self-defining Text Archive and Retrieval (STAR) File uses standard ASCII text to specify both the data structure and the information. The syntax of this file is simple, and it may be easily interpreted visually or by computer. The STAR format is the basis for the Crystallographic Information File (CIF), which has been adopted by the International Union of Crystallography for the submission of data and text to crystallographic journals and data bases.

INTRODUCTION

Many existing computer-archiving procedures use a "fixed format" data structure targeted at specific applications. This approach provides for efficient data access but is inflexible and cannot be changed without reformating existing archived files. These files are unsuitable for long-term archiving of most scientific data where there is a continual evolution of data types.

Another archiving approach is based on "pre-defined free formats". Such formats do not restrict data to specific positions in the file. Often "data keys" are included to aid in data recognition. Examples of this type are the BCCAB archive file[1] used by the Cambridge Crystallographic Data Centre, the Standard Crystallographic File Structure,[2] the JCAMP-DX File[3] for archiving infrared spectra, and the Standard Molecular Data (SMD) Format,[4] a collaborative development of chemical and pharmaceutical laboratories for the global exchange of molecular data. These files, while differing significantly in construction and style, have a common disadvantage: their data syntax is relatively complex and requires careful predefinition to facilitate data access.

The complexity and inflexibility of existing archive files limits the rapid exchange of data, even within disciplines where data requirements are similar. This is a special problem for applications with a continual need for new data items. Some computer-intensive disciplines, such as crystallography, currently support a vast repertoire of specialized and "local" file formats. This was tolerable when electronic data exchange was infrequent and computing considerations required file formats to be finely tuned to specific applications. However, the recent explosion in computer and network performance has signaled an end to this rationale. In an era of increasing global data exchange there is a critical need for a simple but universal archive file.

Table I. A Universal Archive File

·Is used to store *all types* of data
·Is *not necessarily* a data-base file
·Should be *machine independent*
·Should be *simple* to read and to access
·Should be *flexible* to future change

The prerequisites for a universal archive format are simplicity, generality, upwards compatibility, and flexibility (see Table I). Such a file must be machine-independent and portable so that the accessibility of data items is independent of their point of origin. It is fundamental that this file allows data to be incorporated in the future without imposing a need to modify existing files. The Self-defining Text Archive and Retrieval (STAR) format described in this paper is designed to meet these requirements. The STAR file structure is suitable for archiving all types of text and numerical data, in

0095-2338/91/1631-0326$02.50/0

Table II. Properties of a STAR File

- the data *structure* is completely *self-defined*
- the data *items* are completely *self-defined*
- the data *syntax* rules are few and *simple*
- the data may be of *any type* and in *any order*
- the file is *easy to read* visually, or by machine

any order. It is particularly well suited to the requirements of electronic transmission.

PROPERTIES OF A STAR FILE

In this section the general concept of a STAR File is introduced, and its basic properties are described. In the next section precise specifications of the STAR File are given.

The basic properties of a STAR File are summarized in Table II. The STAR File contains textual data that can be edited and read with a standard editor. The text is visually intelligible and can be stored or transmitted electronically without conversion. The syntax of a STAR File is simple. Each file is divided into a sequence of *data blocks* which contain individual data items. The identity of each *data item* is determined by a preceding *data name*. It is possible to repeat data items by placing them within simple looping structures.

Here are some examples of STAR syntax. A *data block* is identified by a unique string, referred to as the *block code*, which is concatenated with 'data_'. The following statement in a STAR File specifies the start of the data block 'crystal_5A5'.

```
data_crystal_5A5
```

Each *data item* is identified by a unique *data name*. A data name is a character string which starts with an underline '_'. Three examples of data names, and their associated data items, follow:

```
_cell_volume                          2310(2)
_chemical_formula                     'C23 H36 O7'
_publication_author_address
;    Prof Barry O'Connell
     Department of Chemistry
     University of Kalamazoo
     Michigan     U.S.A.
;
```

The data items above are of "type" numeric, character and text. The STAR File syntax makes no distinction between data types, other than their recognition as text strings bounded by different delimiters (e.g., blanks, single quotes, and semicolons as the first character of a line). The relative order or format of these strings in the file is irrelevant (apart from the requirement that the data name must precede the data item).

A data item, or a group of data items, may be repeated in a list. These "looped" data items are identified by being preceded by 'loop_' string. Here is a simple example of looped data items:

```
loop_
_exptl_crystal_face_h
_exptl_crystal_face_k
_exptl_crystal_face_l
_exptl_crystal_face_distance
_exptl_crystal_face_name
_exptl_crystal_face_description
 0  0 -1      0.012      A        'well formed'
 0  0  1      0.012      B        .
 1  0  0      0.023      C        uneven
-1  0  0      0.027      D        'requires further grinding'
```

Any data item may be looped. The only requirement is that the number of data items in the loop must be a multiple of the number of data names.

These concepts (summarized in Tables II and III) specify the basic STAR syntactical rules. The simplicity of this syntax is a most important property of a STAR File. It contributes to flexibility and provides wide applicability. No assumptions about the order of the data blocks or data items are made, other than the requirement that identifying names be unique. There are no rules regarding the placement of data names or data items within a data block, other than the requirement that the name must precede the item. Access to data in a STAR File is made simply by requesting a specific data name within a specific data block. No prior knowledge is needed about the data type, whether the item is looped, or whether an item exists in the file.

Table III. A Brief Summary of STAR Terms

- a *data name* is what you call a data item, e.g., _absorption_coefficient
- a *data item* is the data itself, e.g., 15.76(5)
- a *data loop* is a list of repeated data, e.g., a list of intensity measurements
- a *save frame* is a collection of the above, which may be referenced by *framecode* within the data block, e.g., data for a specific molecule
- a *data block* is a collection of the above, e.g., a data set for one compound

Table IV. STAR Syntax

- a *text string* is string of characters bounded by a [ ]['][''][;]
- a *data name* is a text string starting with an underline
- a *data item* is a text string not starting with underline, preceded by an identifying data name
- a *data loop* is a list of data names, preceded by 'loop_' and followed by a repeated list of data items
- a *data block* is a collection of data, preceded by 'data_code'
- a *data file* may contain any number of data blocks
- a *data name* **must** *be unique within a data block*

THE SPECIFICATION OF A STAR FILE[5]

A STAR File is a sequential file containing lines of standard visible ASCII characters. It is divided into any number of parts referred to as data blocks. The information within a data block defines the data structure and the data items. All of this information is intelligible as text.

The following syntax rules specify a STAR File. These are summarized in Table IV.

1. A *text string* is defined as either a sequence of nonblank characters, a sequence of characters bounded by matching single or double quotes (i.e., ⟨'⟩ or ⟨"⟩), or a sequence of lines bounded by a semicolon ⟨;⟩ as the first character of a line. A text string must not span more than one line, except if bounded by semicolons.
2. A *data name* is a text string starting with an underline '_'.
3. A *data item* is a text string *not* starting with an underline '_', and preceded by the identifying data name.
4. A *data loop* is a list of data names, followed by a repeated list of data items, and preceded by the text string 'loop_'.
5. A *save frame* is a sequence of data names, data items, and data loops preceded by the text string 'save_framecode' where 'framecode' is a unique identifying code within a data block. A save frame sequence is closed by another save frame command, by the text string 'stop_', or by a data block command.
6. A *data block* is a sequence of data names, data items, data loops, and save frames preceded by the text string 'data_*blockcode*' where '*blockcode*' is a unique identifying code within a STAR File. The data block sequence is closed by another data block command or the end of the STAR File.

Chart I

```
data_P6122

_audit_creation_date              90-05-25
_audit_creation_method            from_xtal_archive_file_using_CIFIO
_audit_update_record

_computing_data_collection         ?
_computing_cell_refinement         ?
_computing_data_reduction         xtal_ADDREF_SORTRF
_computing_structure_solution     xtal
_computing_structure_refinement   xtal_CRYLSQ
_computing_publication_material   xtal_BONDLA_CIFIO

_cell_a                           8.53(1)
_cell_b                           8.53(1)
_cell_c                           20.37(1)
_cell_alpha                       90.00(1)
_cell_beta                        90.00(1)
_cell_gamma                       120.00(1)
_cell_volume                      1284(1)
_cell_formula_units_Z             24
_cell_measurement_temperature     20
_cell_measurement_reflns_used      ?
_cell_measurement_theta_min        ?
_cell_measurement_theta_max        ?

_symmetry_crystal_system          hexagonal
_symmetry_space_group_name_H-M     ?
_symmetry_space_group_name_Hall   P_61_2___(0_0_-1)
loop_
_symmetry_equiv_pos_as_xyz
  +x,+y,+z  -x,-y,1/2+z  -y,-x,5/6-z  +y,+x,1/3-z  +x-y,-y,-z  -x+y,+y,1/2-z
  +x,+x-y,1/6-z  -x,-x+y,2/3-z  -y,+x-y,1/3+z  +y,-x+y,5/6+z  +x-y,+x,1/6+z
  -x+y,-x,2/3+z

_diffrn_temperature               20
_diffrn_radiation_wavelength       1.5418
_diffrn_radiation_type             ?
_diffrn_radiation_source          xray_tube
_diffrn_radiation_monochromator    ?
_diffrn_radiation_detector         ?
_diffrn_measurement_device         ?
_diffrn_measurement_method         ?

_diffrn_reflns_number             92
_diffrn_reflns_av_R_equivalents   0
_diffrn_reflns_av_sigmaI/Inet     .09408
_diffrn_reflns_h_min              0
_diffrn_reflns_h_max              2
_diffrn_reflns_k_min              0
_diffrn_reflns_k_max              4
_diffrn_reflns_l_min              0
_diffrn_reflns_l_max              12
_diffrn_reflns_theta_min           5.9901
_diffrn_reflns_theta_max          28.9584
_diffrn_reflns_reduction_process   ?

loop_
_atom_type_symbol
_atom_type_oxidation_number
_atom_type_number_in_cell
_atom_type_scat_dispersion_real
_atom_type_scat_dispersion_imag
_atom_type_scat_source
  S  0  6  .319  .557  Int_Tab_Vol_III_p202_Tab._3.3.1a
  O  0  6  .047  .032  Cromer,D.T._&_Mann,J.B._1968_AC_A24,321.
  C  0  12  .017  .009  Cromer,D.T._&_Mann,J.B._1968_AC_A24,321.

loop_
_atom_site_label
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
```

Chart I (*Continued*)

```
_atom_site_U_iso_or_equiv
_atom_site_thermal_motion_type
_atom_site_calc_flag
_atom_site_calc_attached_atom
_atom_site_type_symbol
  s   .2015(4)  -.2015(4)  .91667  .030(6)  Uij  ?  ?  s
  o   .500(3)   .500(3)   .66667  .08(2)   Uij  ?  ?  o
  c1  .492(4)   .096(3)   .037(1)  .03170  Uij  ?  ?  c

loop_
_atom_site_aniso_label
_atom_site_aniso_U_11
_atom_site_aniso_U_22
_atom_site_aniso_U_33
_atom_site_aniso_U_12
_atom_site_aniso_U_13
_atom_site_aniso_U_23
_atom_site_aniso_type_symbol
  s   .041(8)  .041(8)  .012(7)  .024(7)  -.003(6)  -.003(6)  s
  o   .09(2)   .09(2)   .10(2)   .06(2)   .007(8)   -.007(8)  o
  c1  .03170  .03170  .03170  .01585  .00000  .00000  c

_refine_special_details            sfls:_F__unit_weight__full_matrix
_refine_ls_structure_factor_coef   F
_refine_ls_matrix_type             full
_refine_ls_weighting_scheme        unit
_refine_ls_hydrogen_treatment      ?
_refine_ls_extinction_method       Zachariasen_Gaussian
_refine_ls_extinction_coef         .698653
_refine_ls_abs_structure_details   none
_refine_ls_abs_structure_Flack     0
_refine_ls_number_reflns           81
_refine_ls_number_parameters       16
_refine_ls_number_restraints       0
_refine_ls_number_constraints      0
_refine_ls_R_factor_all            .053798
_refine_ls_R_factor_obs            .044049
_refine_ls_R2_factor_obs           .044049
_refine_ls_R2_factor_all           .053798
_refine_ls_wR2_factor_all          .063506
_refine_ls_wR2_factor_obs          .052326
_refine_ls_goodness_of_fit_all     1.8858
_refine_ls_goodness_of_fit_obs     1.6744
_refine_ls_shift/esd_max           1.5163
_refine_ls_shift/esd_mean          .279164
_refine_difference_density_min     -.163089
_refine_difference_density_max     .172854

_geom_special_details              ?
loop_
_geom_bond_atom_site_label_1
_geom_bond_atom_site_label_2
_geom_bond_distance
_geom_bond_site_symmetry_1
_geom_bond_site_symmetry_2
_geom_bond_publ_flag
  s  c1  1.71(3)  666_6  665_2  ?
  s  c1  1.71(4)  666_6  665_8  yes
  o  c1  1.15(4)  566_7  665_2  ?
  o  c1  1.15(3)  566_7  665_6  yes
loop_
_geom_angle_atom_site_label_1
_geom_angle_atom_site_label_2
_geom_angle_atom_site_label_3
_geom_angle
_geom_angle_site_symmetry_1
_geom_angle_site_symmetry_2
_geom_angle_site_symmetry_3
_geom_angle_publ_flag
  c1  s  c1  135(1)  665_2  666_6  665_8  yes
  c1  o  c1  130(1)  665_2  566_7  665_6  yes
  s  c1  o  160(1)  666_6  665_2  566_7  yes
```

Chart II

```
data_P6122

_hist
;
STARTX 25/ 5/90 10:14:27ADDREF 25/ 5/90 10:14:29SORTRF 25/ 5/90 10:14:30
ADDATM 25/ 5/90 10:14:32FC     25/ 5/90 10:14:33CRYLSQ 25/ 5/90 10:14:36
BONDLA 25/ 5/90 13: 3:44BONDLA 31/ 5/90 23:46:44FOURR  31/ 5/90 23:46:45
;
_labl
;
25/ 5/90 10:14:27
 Test case from Larson -- dummy P6122 structure.
;
loop_
    _cell_pak_1
  .853000+01  .853000+01  .203700+02  .000000+00  .000000+00 -.500000+00
  .250000+00  .250000+00  .333333+00

loop_
    _cell_pak_2
  .100000-01  .100000-01  .100000-01  .174533-03  .174533-03  .151150-03
  .277778-04  .277778-04  .277778-04

loop_
    _cell_pak_3
  .135369+00  .135369+00  .490918-01  .000000+00  .000000+00  .500000+00
  .250000+00  .250000+00  .166667+00

_symm_pak_1_1_lattice_type                  1
_symm_pak_1_2_centro_type                   1
_symm_pak_1_3_total_symops                 12
_symm_pak_1_4_basis_symops                 12
_symm_pak_1_5_equiv_symops                  2
_symm_pak_1_6_multiplicity                  1
_symm_pak_1_7_cedar_symops                  0
_symm_pak_1_8_moles/cell                   24
_symm_pak_1_9_cryst_system                  7
_symm_pak_1_10                              0
_symm_pak_1_11                              0
_symm_pak_1_12                              0

loop_
    _symm_pak_r11  _symm_pak_r21  _symm_pak_r31  _symm_pak_r12  _symm_pak_r22
    _symm_pak_r32  _symm_pak_r13  _symm_pak_r23  _symm_pak_r33  _symm_pak_t1
    _symm_pak_t2   _symm_pak_t3
    1    0    0    0    1    0    0    0    1   .000000   .000000   .000000
   -1    0    0    0   -1    0    0    0    1   .000000   .000000   .500000
    0   -1    0   -1    0    0    0    0   -1   .000000   .000000   .833333
    0    1    0    1    0    0    0    0   -1   .000000   .000000   .333333
    1    0    0   -1   -1    0    0    0   -1   .000000   .000000   .000000
   -1    0    0    1    1    0    0    0   -1   .000000   .000000   .500000
    1    1    0    0   -1    0    0    0   -1   .000000   .000000   .166667
   -1   -1    0    0    1    0    0    0   -1   .000000   .000000   .666667
    0    1    0   -1   -1    0    0    0    1   .000000   .000000   .333333
    0   -1    0    1    1    0    0    0    1   .000000   .000000   .833333
    1    1    0   -1    0    0    0    0    1   .000000   .000000   .166667
   -1   -1    0    1    0    0    0    0    1   .000000   .000000   .666667

_sgnm_pak_1
   P_61_2___(0_0_-1)

loop_
    _ddef
 'PARENT      '

loop_
    _atom_0014 _atom_0017 _atom_0021 _atom_0022 _atom_0001 _atom_0002
    _atom_0003 _atom_0004 _atom_0005 _atom_0006 _atom_0007 _atom_0008
    _atom_0009 _atom_0010 _atom_0011 _atom_0023 _atom_0101 _atom_0102
    _atom_0103 _atom_0104 _atom_0105 _atom_0106 _atom_0107 _atom_0108
    _atom_0109 _atom_0110 _atom_0111
```

Chart II (*Continued*)

```
 's'    1  .500000    1  .201493 -.201493  .916670  .030046  .041490  .041490
  .011877  .024286 -.266471-02 -.266471-02    1    2  .436975-03  .436975-03
    0  .627330-02  .788800-02  .788800-02  .720579-02  .706542-02  .603786-02
  .603786-02    0
 'o'    1  .500000    2  .500101  .500101  .666670  .084444  .089157  .089157
  .099369  .062840  .671791-02 -.671791-02    1    2  .272479-02  .272479-02
    0  .015001  .018808  .018808  .021515  .020001  .803740-02  .803740-02   0
 'c1'    1    1    3  .491510  .096455  .037280  .031700  .031700  .031700
  .031700  .015850    0    0  .974722    2  .362040-02  .278182-02  .104156-02
    0    0    0    0    0    0    0  .049621

loop_
    _cons_0001 _cons_0011 _cons_0002 _cons_0005 _cons_0012 _cons_0013
    _cons_0003 _cons_0004
 's' 's'    2    1    1   -1    1    0
 's' 's'    3    3    1    0    1  .916670
 's' 's'    5    1    4    1    1    0
 's' 's'    9    1    8    1    1    0
 'o' 'o'    2    1    1    1    1    0
 'o' 'o'    3    3    1    0    1  .666670
 'o' 'o'    5    1    4    1    1    0
 'o' 'o'    9    1    8   -1    1    0

loop_
    _refl_0001 _refl_0002 _refl_0003 _refl_1600 _refl_1308 _refl_1304
    _refl_1305 _refl_1800 _refl_1801 _refl_1802 _refl_1803 _refl_1804
    _refl_1805 _refl_1806 _refl_1701 _refl_1700 _refl_1202
'   0   0   6'  .147275 Z000010c1  .410000    1  39.4000   3.5100  35.1222
 -35.1222    0  -1.6066  -3.0963 -36.7288  -3.0964  39.3109  .500001  .909983
'   0   0  12'  .294551 Z000010c1  .900000    1  48.4400   2.3800  39.9100
  39.9100  .268379-04   1.7559   3.0979  41.6658   3.0980  47.8384    0
.939517
..... data removed here
'   2   3   2'  .299087 Z0000002c   1.0600    1  22.9200   2.3100  18.8304
  14.1504 -12.4238   1.1726  .012972  15.3230 -12.4108  22.0697  .885327
  .985649
'   2   3   3'  .304082 Z0000002c   1.9100    1  40.3600   2.1900  34.1136
 -33.6726  -5.4677  -1.0348  -1.8594 -34.7074  -7.3271  39.7222  .525621
  .956049
'   2   3   4'  .310939 Z0000002c  .820000    1  16.9200   1.9100  14.0198
  -6.6982 -12.3162  .626657 -.867826  -6.0716 -13.1841  16.4194  .670724
  .992303

loop_
    _sfls_0001 _sfls_0011 _sfls_0012 _sfls_0013 _sfls_0014 _sfls_0015
    _sfls_0016 _sfls_0021 _sfls_0022 _sfls_0023 _sfls_0024 _sfls_0025
    _sfls_0026 _sfls_0027 _sfls_0028 _sfls_0029 _sfls_0030 _sfls_0031
    _sfls_0032 _sfls_0033 _sfls_0041 _sfls_0042 _sfls_0043 _sfls_0044
    _sfls_0045
    1 'sfls: F     unit weight      full matrix' ' unit'
'   Zachariasen Gaussian' '     full' '' 'none'  .053798  .044049  .053798
  .044049  .063506  .052326   1.8858   1.6744   1.5163  .279164    0    0
  .698653    1   16   81    0    0

loop_
    _dens_0001  _dens_0011  _dens_0012
    1 -.162752  .173522

loop_

    _bond_0011 _bond_0012 _bond_0013 _bond_0014 _bond_0015 _bond_0016
    _bond_0021 _bond_0022 _bond_0023 _bond_0024 _bond_0025 _bond_0026
    _bond_0091 _bond_0092
 's'  .597015  .798508  .583328    6   666.00 'c1'  .508492  .903542  .537277
    2   665.00   1.7115  .030558
 's'  .597015  .798508  .583328    6   666.00 'c1'  .508492  .604950  .629389
    8   665.00   1.7117  .042398
 'o'  .500099    1  .500000    7   566.00 'c1'  .508492  .903542  .537277    2
   665.00   1.1479  .044565
 'o'  .500099    1  .500000    7   566.00 'c1'  .604950   1.0965  .462723    6
   665.00   1.1479  .033356
```

7. A data name must be unique within each save frame or, if there are no save frames, within each data block. A save frame declaration must be unique within a data block sequence. The save frame code may be referred to within a data block as the data item '*$framecode*'.
8. Except if contained within a text string, a sequence of blank or tab characters is used only to separate text strings.
9. Except if contained within a text string, a single sharp '#' signals that the characters up to the end of the line are used for comment only.

These specifications define the STAR File syntax completely. There are no restrictions on the order that data is stored, and only minor constraints on the format of data (see specification 1). The only information required to access a specific data item is its *data name*.

To facilitate access to a STAR File the names of data items must be defined to be as descriptive as possible. This is particularly important if a file is to be used for "global" applications (i.e., outside of the local environment where it was generated). If a file is generated for archiving it is fundamental that the data names and their definitions not be changed in the lifetime of the file. Any redefinitions would, in effect, make these data items inaccessible. New name definitions and data items may, however, be added as needed. Because there are no relational constraints between data names, local and global data items may also be mixed freely without prejudicing access to either.

SPECIFIC APPLICATIONS OF A STAR FILE

The versatility of the STAR File format is best demonstrated by actual applications. The first example is the use of the STAR format in the development of the Crystallographic Information File (CIF), recently adopted by the International Union of Crystallography for the submission of data to journals and data bases.[6] In the second application the STAR format is used to port data from the Xtal3.0 Program Package[7] between different sites and machines.

The IUCr CIF Application. A working party was set up by the International Union of Crystallography in 1987 to coordinate the IUCr publishing and data-base activities. A principal objective was the design of a file suitable for the global exchange of crystallographic data. Because crystallographic techniques have many scientific and industrial applications in fields ranging from engineering to medicine, this file had to accommodate a wide spectrum of data items and be capable of change as necessary. The STAR File was selected by the working party as the most appropriate format for this application, and the developed file is referred to as the Crystallographic Information File.

Extracts from a CIF are shown in Appendix I. Each data item in this CIF is identified by a unique data name defined in a CIF Dictionary[8] that contains those data names currently accepted for submission to IUCr journals and crystallographic data bases. A CIF may contain *any* data item, but only those defined in the CIF Dictionary will be recognized globally without further explanation. The inclusion of nonstandard names and data in no way affects the logical integrity of a CIF, or prevents access to the standard data.

Additional syntax restrictions were imposed on the CIF to simplify software development and expedite implementation. The restrictions imposed by the CIF developers,[6] which could be applied in other scientific applications, are listed here.

1. Lines may not exceed 80 characters in length.
2. *Data names* and *block codes* may not exceed 32 characters in length.
3. A data item is assumed to be of type *number* if it starts with either a digit '0'–'9', a plus '+', a minus '–', or a period '.' and is *not* bounded by matching single or double quotes. A number may be in integer, real, or scientific format. If a number is concatenated with another number bounded by parentheses, it is taken to be the standard deviation [e.g., nn.nnn(m)].
4. The data type is assumed to be *character* if it is surrounded by matching single and double quotes, or if it is neither type *number* nor type *text*.
5. Only one level of 'loop_' data is permitted. Additional levels of repeated data must be stored as lists within single text strings.
6. The *save frame* command is not invoked. This simplifies parsing because each data name is globally unique within a data block.
7. Most numeric fields contain data for which the units must be known. Each CIF data item has default units specified in a CIF Dictionary. If nonstandard units are used, a units code is appended to the data name. For example, the standard units for cell dimensions are angstroms. If this data item is included in a CIF with the units of picometers, the standard data name of '_cell_a' would be '_cell_a_pm'.

An extract from a CIF generated within the Xtal3.0 package by the program CIFIO[9] is shown in Appendix I. This CIF is for transmission to a journal and contains only those items requested by the user. Requested data items not present in the Xtal archive files are flagged with a question mark '?'. Such question marks are important for two reasons. They indicate that data items are unavailable, and they maintain the STAR requirement that every data name must be matched with a data value. The missing items can be filled in manually using a text editor. Most manuscript items are added in this way. In some cases the manuscript text and diagrams will be added as *PostScript or TEX or TIFF data*.

It must be emphasized that a CIF need not contain every data item listed in the CIF Dictionary. All data items are optional and should be included only if required. While the order for data items in a CIF is not fixed, certain data items should, for convenience, appear in a common list or loop structure. It is always good practice to group data items according to data category (e.g., all absorption data would be entered together), simplifying the task of manual searching and editing.

The STAR format permits *any* data item to be included in a 'loop_' structure. This only occurs, however, if a data item is repeated in a "list". Clearly data items should appear in the *same* list if they belong to the same basis set, while others in a different basis set will appear in a separate list. For example, '_atom_' data items will form one of these looped lists, while '_refln_' data will form another, and they will not be mixed.

The Xtal Archive File Application. The second application emphasizes the diversity of styles that the STAR format provides. The Xtal3.0 package currently uses the STAR format to archive a complete set of data accumulated on binary files from crystallographic calculations. An extract from one of these archive files, shown in Appendix II, contrasts strongly with the CIF from the same data, shown in Appendix I.

The Xtal archive file is a specialized "local" STAR application. It is not intended for global use, and its data names are not CIF standard names. Nevertheless, it is completely self-contained and portable. It may be transferred from machine to machine, or to another laboratory, and used to produce identical binary files, which can be used for further Xtal calculations.

The Xtal archive file, in STAR format, may also be edited, either manually or with other software (see QUASAR below).

This provides previously unavailable facilities for specialized data updating and editing, and easy access to Xtal data by other program systems.

ACCESSING DATA IN A STAR FILE

Data in a STAR File may be easily read and manipulated, manually or computationally. Access to a specific data item depends only on locating the appropriate data name in the file. No a priori knowledge of the file structure is required. This means that STAR file search procedures can use simple parsing algorithms to access data.

The program QUASAR[10] uses a straightforward parsing approach to retrieving data from a STAR File according to a "request list" of data names. Requested data items and data blocks are output, as another STAR File, in the order requested. Within each data block the same data item may be requested multiply, and "wild card" requests are permitted. QUASAR checks for the logical integrity of the data names and items in a file, and for correct loop counts.

To illustrate the QUASAR approach, here is an example run where data items are extracted from the file *p6122.cif* shown in Appendix I. The QUASAR request list follows:

```
star_arc_p6122.cif
star_out_p6122.out

data_P6122

_audit_creation_date
_audit_creation_method
_audit_update_record
_symmetry_equiv_pos_as_xyz
_geom_contact_atom_site_label_1
_geom_contact_atom_site_label_2
_geom_contact_distance
_geom_contact_site_symmetry_1
_geom_contact_site_symmetry_2
_geom_contact_publ_flag
```

Note that the file names of the input and output STAR Files are specified in the request list as extensions to 'star_arc_' and 'star_out_', respectively. In this example these files are named *p6122.cif* and *p6122.out*.

The output file *p6122.out* generated by the above requests is shown below. Compare this with the data items in *p6122.cif* shown in Appendix I. Note that the output file is also a STAR File.

```
data_P6122

_audit_creation_date              90-05-25
_audit_creation_method            from_xtal_archive_file_using_CIFIO
_audit_update_record              '90-05-26 _geom publ flags added manually'

loop_
_symmetry_equiv_pos_as_xyz
       +x,+y,+z
    -x,-y,1/2+z
    -y,-x,5/6-z
    +y,+x,1/3-z
     +x-y,-y,-z
  -x+y,+y,1/2-z
  +x,+x-y,1/6-z
  -x,-x+y,2/3-z
  -y,+x-y,1/3+z
  +y,-x+y,5/6+z
  +x-y,+x,1/6+z
  -x+y,-x,2/3+z

loop_
_geom_bond_atom_site_label_1
_geom_bond_atom_site_label_2
_geom_bond_distance
_geom_bond_site_symmetry_1
_geom_bond_site_symmetry_2
_geom_bond_publ_flag
  s  c1  1.71(3)  666_6  665_2    ?
  s  c1  1.71(4)  666_6  665_8  yes
  o  c1  1.15(4)  566_7  665_2    ?
  o  c1  1.15(3)  566_7  665_6  yes
#                 -----end-of-data-block-----
```

APPENDIX I. EXAMPLE 1 OF A STAR FILE: EXTRACT FROM A 'PUBLICATION' CIF

This (Chart I) is an extract from a CIF generated by the program CIFIO[9] using structural data 'P6122' from an Xtal binary file. This style of STAR File is intended for publication purposes and contains only data items requested by the user. Items not present in the Xtal binary file are flagged with a '?'.

APPENDIX II. EXAMPLE 2 OF A STAR FILE: EXTRACT FROM AN XTAL ARCHIVE FILE

This (Chart II) is an extract from a Xtal archive file generated by the program CIFIO[9] using the structural data set 'P6122' stored on an Xtal3.0 binary file. This style of STAR File is intended for archival purposes or data communication with other Xtal users.

ACKNOWLEDGMENT

The STAR File concepts were first proposed by the author in 1978 for an archive file project of the IUCr Commissions on Crystallographic Computing and Data.[11] At that time they were considered too advanced for the then-current level of software development. These concepts were resurrected in 1987 in response to a need to store a range of diffraction data generated at different sites. This led to the first set of STAR specifications, and an initial version of the QUASAR software. Some refinement of these concepts took place over the next 2 years following the selection of the STAR File as the basis for the CIF development. For the advice, assistance, discussions, and challenges offered by the following colleagues, Frank Allen, David Brown, Mike Dacombe, Howard Flack, Richard Goddard, Carl Krüger, Ted Maslen, Brian McMahon, George Sheldrick, Rolf Sievers, and Jim Stewart, I am indebted. Thanks are also due to the 1987–90 IUCr Executive and the Working Party on Crystallographic Information for their strong support of the CIF project which was the catalyst for many the STAR File refinements.

REFERENCES AND NOTES

(1) The BCCAB archive file is used by the Cambridge Data Centre (U.K.) to prepare the packed organic structural data-base file ASER.

(2) Brown, I. D. The Standard Crystallographic File Structure. *Acta Crystallogr.* **1988**, *A44*, 232.

(3) McDonald, R. S.; Wilks, P. A., Jr. JCAMP-DX: A Standard Form for Exchange of Infrared Spectra in Computer Readable Form. *Appl. Spectrosc.* **1988**, *43*, 151–162.

(4) Barnard, J. N. Draft Specification for Revised Version of the Standard Molecular Data (SMD) Format. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 81–96.

(5) Copyright of the STAR specifications is held by the International Union for Crystallography, 5 Abbey Square, Chester, U.K., and a patent has been applied for to cover the STAR File process.

(6) Hall, S. R.; Allen, F. H.; Brown, I. D. The Specification of the CIF Standard Data. *Acta Crystallogr.* In preparation.

(7) Hall, S. R.; Stewart, J. M. Xtal3.0 Crystallographic Program System. Publication of the University of Western Australia and University of Maryland, 1990.

(8) Hall, S. R.; Allen, F. H.; Brown, I. D. The CIF Dictionary. *Acta Crystallogr.* In preparation. Copies of the dictionary may be obtained from the IUCr, 5 Abbey Square, Chester CH1 2HU, England.

(9) Hall, S. R. CIFIO: Xtal3.0 Crystallographic Program System. Hall, S. R., Stewart J. M., Eds.; Publication of the University of Western Australia and University of Maryland, 1990.

(10) Hall, S. R.; Sievers, R. QUASAR: A Program for Accessing a STAR File. A copy of this program may be obtained from the authors free of charge.

(11) Brown, I. D. The Standard Crystallographic File Structure. *Acta Crystallogr.* **1983**, *A39*, 216–224.