Semantically rich metadata in crystallographic publishing

Brian McMahon

International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK
bm@iucr.org

Abstract

Semantically-rich metadata assists publisher workflow and greatly increases the added value to the reader of a scientific publication. Where once bibliographic metadata provided the sole opportunity to locate and retrieve scientific publications, the experimental data on which scientific results are based itself is rich in metadata characterizing the data and the results derived from it. In crystallography, a community data exchange standard exists that has been used by specialist journals for over a decade and is increasingly enabling scientists to populate open-archive repositories of experimental results. By extracting and harvesting the relevant metadata, the possibility arises of linking data repository, journal and structure database together in a networked, distributed information resource. The metadata also have an essential role to play in assuring the quality of scientific data and publications.

Keywords: metadata workflow crystallography

1. Historical generation and management of bibliographic metadata

For many years the creation and management of bibliographic metadata – information about published articles – was the preserve of abstracting and indexing services or secondary content aggregators. Scientific publishers were content to create the printed journal, and leave it to organizations such as ISI, INSPEC and *Chemical Abstracts* to create collections of contents listings, citations or abstracts. In some cases the entire contents were read by specialist abstractors who would index relevant scientific information. The immense success of the *Chemical Abstracts* Service for almost a century is a testimony to the value of these aggregators and indexers. For many years, the International Union of Crystallography (IUCr) produced and sold in respectable numbers annual hard-bound volumes containing abstracted summaries of crystal and molecular structures gleaned from its own and other journals.

2. Metadata generation in-house

With the increase in computerized automation of publishing procedures towards the end of the last century, the primary publishers began to find themselves in possession of growing volumes of bibliographic and other metadata in electronic format. There was a clear incentive to make use of this information to organize internal workflows, to provide more efficient supply of metadata to secondary publishers, and gradually to provide value-added services directly to the end-user, the reader.

3. Case study

In the early 1990s the IUCr already had a long-standing practice of creating cumulative (hard-copy) indexes of its journal contents. As computers were introduced into the journal office, in-house composition of these indexes was undertaken to save money and time, but it rapidly became apparent to us that programmatic processing of the embedded typographic markup in the contents list could easily generate moderately well structured metadata records. This technique was used to create electronic indexes of contents (for distribution on floppy disk). Shortly afterwards, we were encouraged to conduct our first experiments in Internet delivery by the appearance of early distributed information protocols such as *gopher* [3] and *WAIS* [13]. *Gopher* was a general mechanism to access material categorized and published from a computerized information store, and was replaced as a delivery mechanism within a few years by the http protocol that underlies today's World Wide Web.

WAIS, however, was in some ways more interesting. It was a wide-area information server application designed to service queries conforming to the Z39.50 information retrieval protocol [4], and provided an effective and lightweight distributed search engine. It possessed a full-text search capability, which we used to provide a query interface for article titles (at that time, of course, the full text of the articles was not available to us). We are well aware how full-text search engines on the Internet have flourished in the last decade, but even large-scale projects such as Google are showing signs of reaching the limits of their usefulness, and there has recently been fresh movement towards promoting distributed full-text query systems. The OpenSearch protocol for query syndication [1] is perhaps an exciting foretaste of what is to come.

Although *WAIS* is now defunct, there have been recent developments in improving the functionality of Z39.50-based distributed query applications, such as the Common Query Language CQL (http://www.loc.gov/z3950/agency/zing/cql/)

and the Search and Retrieve Web and URL Services SRW/U (http://www.loc.gov/z3950/agency/zing/srw/) of the Library of Congress, although it is noteworthy that no large-scale open distributed databases of bibliographic information are yet available. There is limited public access through a query interface to the metadata holdings of the linking aggregator CrossRef (http://www.crossref.org/guestquery/), but this is not complete and is not widely known. There is also no guarantee that such an open service will continue indefinitely under the aegis of a commercial organization. These may be some of the reasons why community-based syndication of bibliographic references through novel services such as Connotea (www.connotea.org) and CiteULike (www.citeulike.org) is causing a stir at the moment.

We shall return to the topic of distributed querying later in this paper.

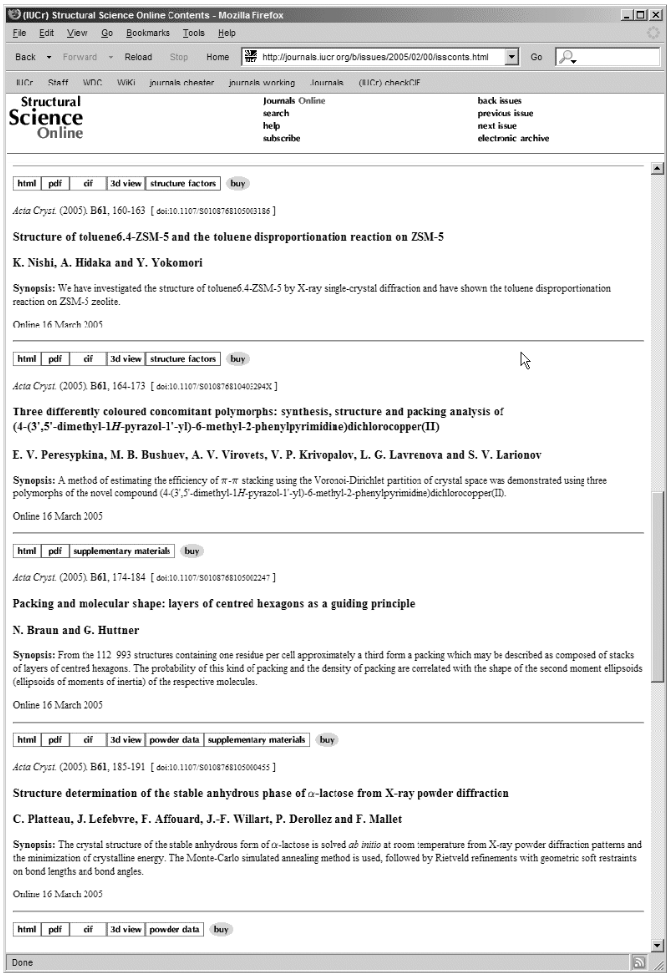
An interesting corollary of the IUCr's early experiments with electronic indexes was that we were able to use the metadata we had accumulated for past publications to create a workflow when we decided to digitize our entire back catalogue of journals. We worked with the United Kingdom Higher Education Digitisation Service (http://heds.herts.ac.uk) on scanning and optical character recognition specifications suitable for creating a PDF archive and searchable full-text index of 50,000 articles in 200,000 pages, over a 50-year publication time span. The project was completed ahead of time and below its modest budget, thanks largely to the availability of structured bibliographic records which we used to create inventories in spreadsheet form, delivery manifests and work schedules.

4. Published Items database

The IUCr decided in 1999 to host its own online publishing operation, and an essential part of the design process involved the commissioning of a 'published items' database that would store all the metadata needed to characterize, locate and deliver any individual article. An early point of debate was how to assign a unique and permanent identifier to published items, the main contenders at the time being the Publisher Item Identifier (PII) of the STI group, and the ANSI/NISO Standard Serial Item and Contribution Identifier (SICI) [5]. The PII had the benefit of being at least in part a 'dumb' identifier – it could incorporate local workflow identification codes through the production and dissemination lifecycle of an article. On the other hand, the SICI had an internal structure allowing a mapping to be made from known bibliographic information about an article to its unique identifying code. In brief, the SICI incorporated known bibliographic metadata, but could only be assigned upon publication; the PII was useful to the publisher in pre-publication workflow, but was unguessable.

We resolved the difficulty by associating both identifiers to each article: the PII was derived from the editorial job ticket, the SICI was assigned upon publication. To handle both legitimate identification handles, we needed an identifier resolver. In retrospect, this architecture was well chosen, since simple extensions of the resolver can now handle additional standard identifiers, such as the Digital Object Identifier (DOI), as well as parsing a variety of query strings based on known bibliographic metadata, whether a full SICI, an openURL representation, or ad hoc URLs mirroring volume/issue/page assignments.

The use of PII as one of the identifiers was also useful in integrating the Published Items database with our production database; as a consequence, individual components of the article (such as figures, tables, and, more importantly, supporting materials and data sets) could be tracked and identified within the Published Items database as the need arose, and had their own individual unique identifiers. From the launch of our online journals service, we were therefore in a good position to provide access to supplementary materials as an integral part of the online publication (Fig. 1).


(IUCr) Structural Science Online Contents - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Back Forward Reload Stop Home
http://journals.iucr org/b/issues/2005/02/00/issconts.html
Go
IUCr Staff WDC WiKi journals chester journals working Journals (IUCr) checkCIF
Structural Science Online
Journals Online
search
help
subscribe
back issues
previous issue
next issue
electronic archive
Done

html | pdf | cif | 3d view | structure factors | buy

Acta Cryst. (2005). **B61**, 160-163 [doi:10.1107/S0108768105003186]

Structure of toluene6.4-ZSM-5 and the toluene disproportionation reaction on ZSM-5

K. Nishi, A. Hidaka and Y. Yokomori

Synopsis: We have investigated the structure of toluene6.4-ZSM-5 by X-ray single-crystal diffraction and have shown the toluene disproportionation reaction on ZSM-5 zeolite.

Online 16 March 2005

html | pdf | cif | 3d view | structure factors | buy

Acta Cryst. (2005). **B61**, 164-173 [doi:10.1107/S010876810403294X]

Three differently coloured concomitant polymorphs: synthesis, structure and packing analysis of (4-(3',5'-dimethyl-1*H*-pyrazol-1'-yl)-6-methyl-2-phenylpyrimidine)dichlorocopper(II)

E. V. Peresypkina, M. B. Bushuev, A. V. Virovets, V. P. Krivopalov, L. G. Lavrenova and S. V. Larionov

Synopsis: A method of estimating the efficiency of π-π stacking using the Voronoi-Dirichlet partition of crystal space was demonstrated using three polymorphs of the novel compound (4-(3',5'-dimethyl-1*H*-pyrazol-1'-yl)-6-methyl-2-phenylpyrimidine)dichlorocopper(II).

Online 16 March 2005

html | pdf | supplementary materials | buy

Acta Cryst. (2005). **B61**, 174-184 [doi:10.1107/S0108768105002247]

Packing and molecular shape: layers of centred hexagons as a guiding principle

N. Braun and G. Huttner

Synopsis: From the 112 993 structures containing one residue per cell approximately a third form a packing which may be described as composed of stacks of layers of centred hexagons. The probability of this kind of packing and the density of packing are correlated with the shape of the second moment ellipsoids (ellipsoids of moments of inertia) of the respective molecules.

Online 16 March 2005

html | pdf | cif | 3d view | powder data | supplementary materials | buy

Acta Cryst. (2005). **B61**, 185-191 [doi:10.1107/S0108768105000455]

Structure determination of the stable anhydrous phase of α-lactose from X-ray powder diffraction

C. Platteau, J. Lefebvre, F. Affouard, J.-F. Willart, P. Derollez and F. Mallet

Synopsis: The crystal structure of the stable anhydrous form of α-lactose is solved *ab initio* at room temperature from X-ray powder diffraction patterns and the minimization of crystalline energy. The Monte-Carlo simulated annealing method is used, followed by Rietveld refinements with geometric soft restraints on bond lengths and bond angles.

Online 16 March 2005

html | pdf | cif | 3d view | powder data | buy

Fig. 1: Extract from the online contents list of a typical issue of *Acta Crystallographica Section B: Structural Science*. The icons above each abstract allow the reader to retrieve the full text of the article (in HTML or PDF formats), or

supplementary information of many types, such as primary experimental data sets, data describing crystal structure models, molecular visualizations, deposited appendices or lengthy tabulations, movies *etc*.

5. Metadata export

An immediate advantage to having a fully populated metadata database is that subsets can be exported in any desired format, so that the transfer of bibliographic information to the traditional abstracting and indexing services becomes very easy, and can be integrated into the production workflow. It is worth noting that at this time, our supply of metadata to secondary publishers and aggregators is a simple step within the production cycle, but is still initiated manually on account of the different procedures that are necessary to deal with a variety of partners. We anticipate that over time recipients of metadata will download fresh metadata as it becomes available; the availability will be made known through e-mail alerting, RSS feed or, perhaps most usefully, through a metadata harvesting protocol such as OAI-PMH [14].

6. Archiving workflow

Another benefit of an efficient metadata publication mechanism such as OAI-PMH is its potential for helping to maintain synchronicity with off-site archives. In common with many publishers, we are exploring collaborative programmes with a number of major national libraries to allow the secure archiving of electronic publications and so guarantee their long-term preservation. Among the procedures being considered for automating ingest by these archives is the use of OAI-PMH to trigger retrieval of new content and to synchronise holdings against any revisions of previously retrieved content.

7. Scientific metadata in crystallography

The preceding discussion may serve as an overview of metadata management that is common to all publishers now working in the electronic environment. Now we consider aspects of scientific publishing that are particularly relevant to data-rich subjects such as crystallography. Crystallography is the branch of science devoted to the study of molecular and crystalline structure, with far-reaching applications in chemistry, physics, mathematics, biology and materials science. X-ray diffraction analysis of crystal structures allows the discovery of how materials pack in the solid state, and of the three-dimensional structure of molecules (including the large biological macromolecules such as DNA and proteins that lie at the heart of structural biology).

Many publications in crystallography include a discussion of aspects of crystal or molecular structure, and indeed a high proportion of scientific articles published in our journals are straightforward reports of new molecular structure determinations. It has become established practice in the discipline to support an article with a deposited data set. This set may contain the final three-dimensional coordinates of the atoms in a molecule and their atomic displacement parameters (a measure of how rigidly bound they are in the crystal lattice), together with various experimental metrics to indicate the quality of the structure determination; or it may include the initial experimental data, so that subsequent researchers can redetermine the structure with different mathematical models. A number of important structural databases collect the coordinate and atomic displacement information from the published literature, and may also accept deposited data referring to unpublished structures. The journals of the IUCr require that every article describing a small-molecule or inorganic structure must be accompanied by the coordinate and displacement data and the initial experimental observations from which they have been derived. Not all journals in the field exercise such a mandatory policy, but most will accept at least the description of the structural model as supporting material. For studies of biological macromolecules, the data sets are large in size, and it is less practical for individual journals to archive such data sets; but the practice in that case is increasingly to require that the author has deposited the data with the Protein Data Bank, which is regarded as the authoritative archive of macromolecular structures.

The relevant Scientific Unions are actively involved in promoting data archival policies among publishers and databases. While there is similar interest and concern with other forms of scientific data, it is probably fair to say that crystallography is one of the disciplines where such policies are most effective.

The potential problem of archiving such data in a uniform format was addressed by the IUCr in 1991, with the commissioning of a standard data exchange format for small-unit-cell structures [11], subsequently extended to biological macromolecular structures [6]. This format, the Crystallographic Information File (CIF), is an efficient text-based tag–value format that can easily be transformed to other formats as required (for example, to XML, which was developed only after CIF was already widespread in the crystallographic world). Fig. 2 is a simple example of how data are identified and presented in a CIF; the resulting file is easy for a human to browse, but, more importantly, can be parsed easily with appropriate software.

An essential factor in the value of CIF as a data exchange standard is the existence of authoritative libraries of recognized data-name tags. These libraries are maintained as electronic files, themselves in CIF format (so that the same parser can be used to read the data file and its associated tag library). The situation is similar to that of a document type definition (DTD) associated with an XML file. However, the CIF tag libraries contain not just lists of recognized tags, but information about the attributes that may be attached to the associated data item. These attributes include basic properties such as data type (string or numeric) and categorization within a taxonomic classification scheme. However, they also contain information about physical units, range of possible values, and relationships to other data items (Fig. 3). There is also a textual definition of the data item to which the tag refers. These libraries contain so much metadata about the data items themselves that they are known as CIF 'dictionaries'.

```
data_HDMB

_chemical_name_systematic
        'N-(2-hydroxy-1,1-dimethylethyl)benzamide'
_chemical_formula_moiety          'C11 H15 N1 O2'
_chemical_formula_weight          193.24
_symmetry_cell_setting            monoclinic
_symmetry_space_group_name_H-M    'P 1 21/n 1'

_exptl_special_details
; The Laue group assignment, the systematic
absences and the centrosymmetry indicated by the
intensity statistics led to assignment of the space
group uniquely as P2~1~/n (No. 14); since
refinement proceeded well, it was adopted.
;

loop_
     _geom_bond_atom_site_label_1
     _geom_bond_atom_site_label_2
     _geom_bond_distance
        O1 C7     1.2446(11)
        O2 C9     1.4148(13)
        O2 H1O2   0.890(16)
```

Fig. 2: Example of crystallographic data in CIF format. Tags beginning with an underscore character identify specific items of data and are defined in central reference documents.

```
data_cell_volume
     _name                         '_cell_volume'
     _category                       cell
     _type                           numb
     _type_conditions                esd
     _enumeration_range              0.0:
     _units                          A^3^
     _units_detail                 'cubic angstroms'
     _definition
;    Cell volume V in angstroms cubed.
     V = a b c [1 - cos^2^(alpha) - cos^2^(beta) -
           cos^2^(gamma) + 2 cos(alpha) cos(beta)
           cos(gamma) ] ^1/2^

     a     = _cell_length_a
     b     = _cell_length_b
     c     = _cell_length_c
     alpha = _cell_angle_alpha
     beta  = _cell_angle_beta
     gamma = _cell_angle_gamma
;
```

Fig. 3: Example of a CIF dictionary definition where the attributes associated with a physical quantity – here, the volume of a crystallographic unit cell – may be detailed.

In practice, data items used universally through crystallography are defined in the official dictionaries managed by a committee of the IUCr known as COMCIFS (Committee for the Maintenance of the CIF Standard). CIF allows for extensibility through local data names that may be defined by individuals for special purposes; there is a limited name space registry to guard against clashes of name. As an archival standard, the definitions in the official dictionaries are fixed and may not change; regular extensions to the dictionary introduce new data names that either describe completely new physical quantities, or that replace erroneous or outmoded data representations.

8. Capturing metadata in the scientific workflow

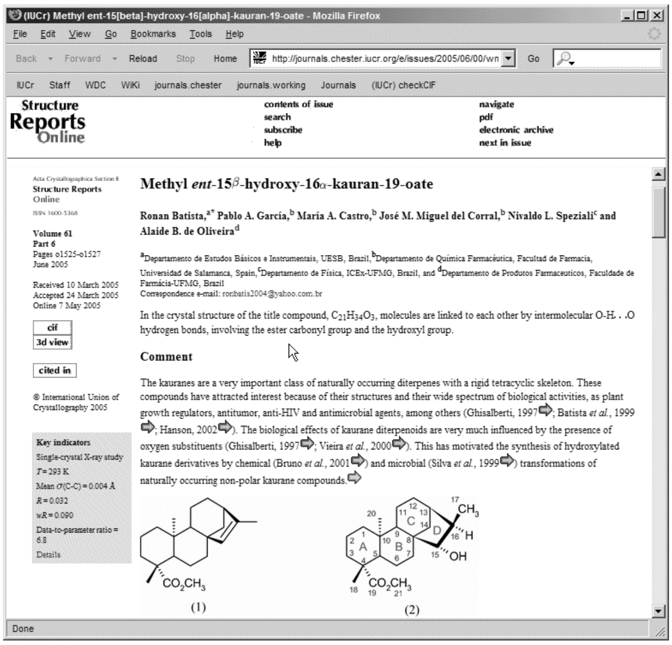
With the existence of community standard dictionaries, it becomes feasible to create information workflows in the collection and processing of experimental data. In the course of a crystallographic experiment, electronic information may be passed from the diffractometer or other data-collecting instrument, through the scientist's data reduction, analysis and modelling software packages on one or more computers, eventually to end up as a file destined for deposition in a database or submission to a journal. It is feasible for this information transfer to involve just a single CIF – created on the diffractometer, populated with the experimental observations, then growing with additional information about the structure as the data are progressively refined and fitted to a model structure. The same file acts as input to, or output from, the variety of software tools that the workbench scientist uses, and eventually may be deposited in a database, or sent as supplementary information to a journal.

The CIF dictionaries currently available [12] define over 3000 data items that list every quantifiable or itemizable aspect of a crystal structure determination in the fields of molecular, inorganic and biological macromolecular single-cell crystallography, powder diffraction, electron density or incommensurate structure studies, and image-plate characterization and measurement. So rich are these compendia, in fact, that some of the IUCr journals use CIF not simply as a container for supplementary data, but as the article submission format itself. That is, from the author's submitted CIF the journal extracts the title of the study, the authors' names and affiliations, an abstract, comment sections and references, as well as the large number of distinct items of experimental and model data that are reported in a standard crystal structure communication. The software used by the journals automatically re-orders this information, builds the necessary tables of experimental detail and molecular geometry, and converts all this information to the SGML used for journal production. This is largely an automated process. While there is some intervention by the technical editor to copy-edit the submission and to iron out small irregularities of presentation, a representation of an article that is already almost suitable for direct publication can be generated without human intervention. One of the services offered to prospective authors is the ability to upload a CIF over the web and receive within seconds a representation of how it will appear in print.

9. Using experimental metadata to characterize structures and quality

From the journal's viewpoint, CIF delivery is valuable in providing important bibliographic metadata (title, authors *etc*.) at the point of submission. However, there is a more dramatic use to be made of the contents of the CIF: it contains the *scientific* metadata needed to characterize the structures reported in the article. Since it began publication as an online-only journal in 2001, *Acta Crystallographica Section E* has extracted from the CIF and presented as part of the article a small set of key indicators, representative experimental metadata useful to the trained crystallographer in summarizing the most relevant properties affecting the experiment and the quality of the structural analysis (Fig. 4).

These key indicators are stored in the Published Items database, although there is as yet no application that accesses them directly. It is clear, however, that they could be used to provide a service for structured querying on the selected experimental criteria. Likewise, such structured queries could be extended to a wider range of metadata elements extracted at will from the submitted CIFs.


(IUCr) Methyl ent-15[beta]-hydroxy-16[alpha]-kauran-19-oate - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Back Forward Reload Stop Home http://journals.chester.iucr.org/e/issues/2005/06/00/wn Go
IUCr Staff WDC WiKi journals.chester journals.working Journals (IUCr) checkCIF
Structure Reports Online
contents of issue
search
subscribe
help
navigate
pdf
electronic archive
next in issue
cif
3d view
cited in
20
11
12
13
C
14
17
CH3
D
16
H
1
2
A
3
4
5
10
9
B
8
6
7
15
OH
18
CO2CH3
19
21
(1)
(2)
Done

Acta Crystallographica Section E
Structure Reports Online
ISSN 1600-5368

Volume 61
Part 6
Pages o1525-o1527
June 2005

Received 10 March 2005
Accepted 24 March 2005
Online 7 May 2005

© International Union of Crystallography 2005

Key indicators
Single-crystal X-ray study
T = 293 K
Mean σ(C-C) = 0.004 Å
R = 0.032
wR = 0.090
Data-to-parameter ratio = 6.8
Details

Methyl *ent*-15β-hydroxy-16α-kauran-19-oate

Ronan Batista,[a*] Pablo A. García,[b] María A. Castro,[b] José M. Miguel del Corral,[b] Nivaldo L. Speziali[c] and Alaide B. de Oliveira[d]

[a]Departamento de Estudos Básicos e Instrumentais, UESB, Brazil, [b]Departamento de Química Farmacéutica, Facultad de Farmacia, Universidad de Salamanca, Spain, [c]Departamento de Física, ICEx-UFMG, Brazil, and [d]Departamento de Produtos Farmaceuticos, Faculdade de Farmácia-UFMG, Brazil
Correspondence e-mail: ronbatis2004@yahoo.com.br

In the crystal structure of the title compound, $C_{21}H_{34}O_3$, molecules are linked to each other by intermolecular O-H. . .O hydrogen bonds, involving the ester carbonyl group and the hydroxyl group.

Comment

The kauranes are a very important class of naturally occurring diterpenes with a rigid tetracyclic skeleton. These compounds have attracted interest because of their structures and their wide spectrum of biological activities, as plant growth regulators, antitumor, anti-HIV and antimicrobial agents, among others (Ghisalberti, 1997⇨; Batista *et al*., 1999 ⇨; Hanson, 2002⇨). The biological effects of kaurane diterpenoids are very much influenced by the presence of oxygen substituents (Ghisalberti, 1997⇨; Vieira *et al*., 2000⇨). This has motivated the synthesis of hydroxylated kaurane derivatives by chemical (Bruno *et al*., 2001⇨) and microbial (Silva *et al*., 1999⇨) transformations of naturally occurring non-polar kaurane compounds.⇨

Fig. 4: Key indicator experimental metadata listed in the header (tinted box on the left-hand side) of a structure communication in *Acta Crystallographica Section E: Structure Reports Online*.

10. Improving scientific quallity

The experimental metadata in a CIF can do more than characterize the compound under study; they can provide an assessment of the quality of the reported structure. Because of the well defined nature of the crystal-structure experiment, reviewers of submitted structure communications have often been able and willing to assess the reported structural model against the experimental data. However, the growth in the number of reported structures during the 1980s made this an increasingly onerous part of the scientific editorial workload, and increasing numbers of errors were beginning to creep into the literature. As a machine-readable format, CIF was ideal for automated extraction and analysis of specific data items and consequent assessment of the reasonableness of the reported structure.

The IUCr now routinely checks submitted structures against a large number of mathematical and logical tests for consistency, scientific reasonableness and data quality (346 distinct tests are currently listed on the IUCr web site at http://journals.iucr.org/services/cif/datavalidation.html).

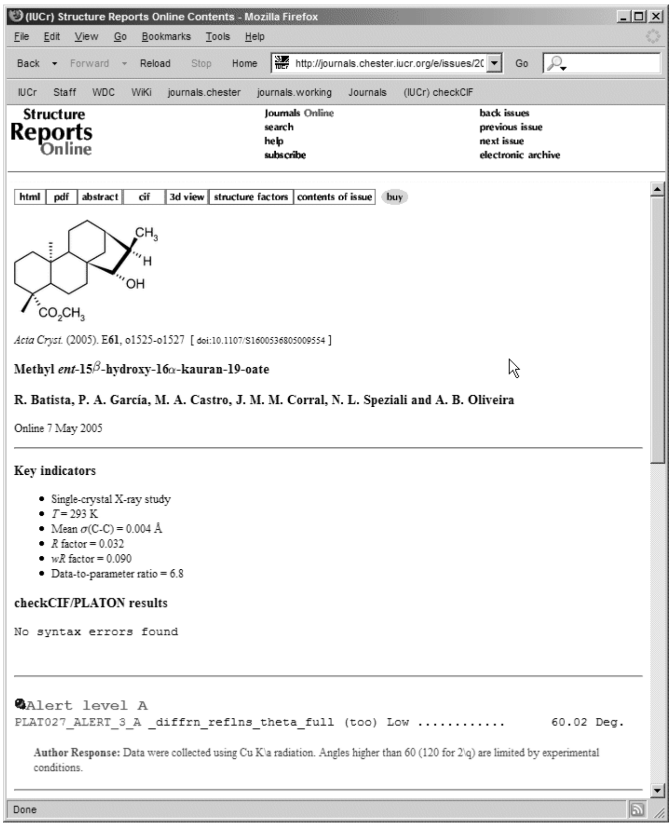
Where values of physical properties are reported that appear to be outliers from the expected norm, the automated reviewing procedure raises a warning flag, and reviewers are expected to check that the author has supplied a reasonable justification for such outlier values. (Articles are not rejected solely on the basis of the automated checks, since there is a small chance that apparently anomalous results are indicative of real, new science. However, indications of a large number of outlier or anomalous values will certainly raise the reviewer's level of scepticism.)

For *Acta Crystallographica Section E*, a summary of the validation report (including a listing of the key indicators and any specific response the author has made regarding outlier values identified in the check report) accompanies every published article as a supplementary document accessible to any reader (Fig. 5).

11. Promoting quality standards

The checking and validation criteria introduced by the IUCr journals are now seen as the 'gold standard' against which structure determinations should be judged, and they have been important in dramatically reducing the number of erroneous structures published. In the 1980s and early 1990s, few issues of the journal did not carry a number of corrigenda to recently published structures – sometimes arising from simple transcription errors in keyboarding data from manuscripts, sometimes reflecting scientific misjudgements. Nowadays, such corrigenda are very uncommon. Transcription errors have been eliminated (the published values are those extracted automatically from the author's working files), and the validation procedures alert reviewers to most potential scientific blunders.

The IUCr now provides a public service (checkcif.iucr.org; see Fig. 6) that allows anyone to check a crystal structure model in CIF format against the checklist. This service is sponsored by several publishers and databases, and used by others to provide support for their own quality assessment procedures. It is not necessarily the case that all the sponsors apply the same quality criteria in selecting articles for publication. Some primarily chemical studies may perform preliminary crystal structure determinations, but not carry them through to the highest achievable level of completeness; or it may be that some materials form very poor quality crystals that yield low-quality results. Nevertheless, used with proper judgement, the *checkcif* tests provide essential information on which to base a considered assessment of the quality of the science reported.


(IUCr) Structure Reports Online Contents - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Back Forward Reload Stop Home
http://journals.chester.iucr.org/e/issues/2(
Go
IUCr Staff WDC WiKi journals.chester journals.working Journals (IUCr) checkCIF
Structure Reports Online
Journals Online
search
help
subscribe
back issues
previous issue
next issue
electronic archive
html pdf abstract cif 3d view structure factors contents of issue buy
CH3
H
OH
CO2CH3
Done

Acta Cryst. (2005). **E61**, o1525-o1527 [doi:10.1107/S1600536805009554]

Methyl *ent*-15β-hydroxy-16α-kauran-19-oate

R. Batista, P. A. García, M. A. Castro, J. M. M. Corral, N. L. Speziali and A. B. Oliveira

Online 7 May 2005

Key indicators

- Single-crystal X-ray study
- T = 293 K
- Mean σ(C-C) = 0.004 Å
- R factor = 0.032
- wR factor = 0.090
- Data-to-parameter ratio = 6.8

checkCIF/PLATON results

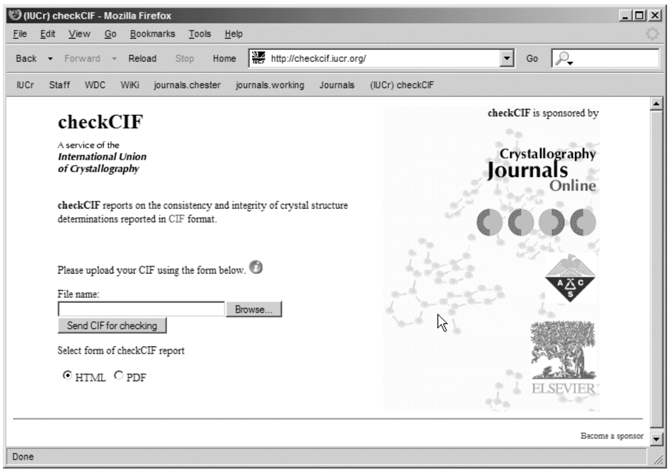
```
No syntax errors found

Alert level A
PLAT027_ALERT_3_A _diffrn_reflns_theta_full (too) Low ...........      60.02 Deg.
```

Author Response: Data were collected using Cu K\a radiation. Angles higher than 60 (120 for 2\q) are limited by experimental conditions.

Fig. 5: Extract from an enhanced *checkcif* report on an article published by *Acta Crystallographica Section E*. Alerts indicating anomalous or outlying values are generated at levels A, B and C, in decreasing order of severity. There is also a class of general alerts (level G) that are not usually of major significance.

12. Access to supplementary data

We have described the way in which our Published Items database allowed us to track individual components of an article. Among these we number the CIF data sets of experimental observations or derived results, and the *checkcif* validation reports. We believe that these are sufficiently important that they should possess globally unique identifiers, allowing them to be located, retrieved and indexed. They are also seen as essential components of the scientific record that should be archived in any long-term preservation project. We are currently assigning individual DOIs to these items within the framework of the CrossRef component model [9].


(IUCr) checkCIF - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Back Forward Reload Stop Home
http://checkcif.iucr.org/
Go
IUCr Staff WDC WiKi journals.chester journals.working Journals (IUCr) checkCIF
Crystallography Journals Online
ELSEVIER
Browse...
Send CIF for checking
Become a sponsor
Done

checkCIF

A service of the *International Union of Crystallography*

checkCIF reports on the consistency and integrity of crystal structure determinations reported in CIF format.

checkCIF is sponsored by

Please upload your CIF using the form below.

File name:

Select form of checkCIF report

● HTML ○ PDF

Fig. 6: Interface to the public *checkcif* service checkcif.iucr.org.

At this stage in the development of the component model, there is not very much provision within the CrossRef schema to describe the nature of individual article components in a structured way. We believe that it would be helpful to have some formal classification scheme that allowed a finely-detailed description of the nature of components representing scientific data sets. The necessary level of detail would include information about the discipline in which the data were collected, the type of experiment that was conducted, and perhaps such information as whether the data set formed a time series or was time-independent.

However, even within the existing schema the ability to assign DOIs to supplemental data sets will greatly enrich their importance as research data available directly from journal holdings.

13. Data repositories

Although there is a rapid and continuing growth in the number of crystal structures being determined (over 24,500 structures were entered in the Cambridge Structural Database in 2001 [2] and over 5000 macromolecular structures were entered in the Protein Data Bank during 2004 [15]), there are growing concerns that not all structure descriptions are finding their way into the public domain. High-volume service crystallography facilities are solving new structures at a rate that makes it difficult to document them fully in a traditional publication. The IUCr launched *Acta Crystallographica Section E* as a relatively low-threshold route to publication, in order to capture as many as possible of the structures flowing out of area-detector-equipped laboratories. The efficient handling of electronic submissions as described above has certainly contributed towards its success (the equivalent of 4663 printed pages was produced in 2004, while submissions during 2005 are already up 65% over the previous year). Nevertheless, there is evidence that even more solved structures are not being submitted for publication in the IUCr or other crystallographic and chemical journals.

Large numbers of structure determinations are being stored in individual laboratories or service crystallography facilities, but many of these institutions have no clear policies for ensuring the long-term preservation of such material. One recent initiative that is beginning to capture the attention of crystallography laboratories is the growth of institutional repositories of publications, typically within universities. With a modest investment in technical infrastructure, the same model can be adapted to the setting up of data repositories, either as components of a sponsoring university's repository, or as a resource managed directly by the laboratory itself.

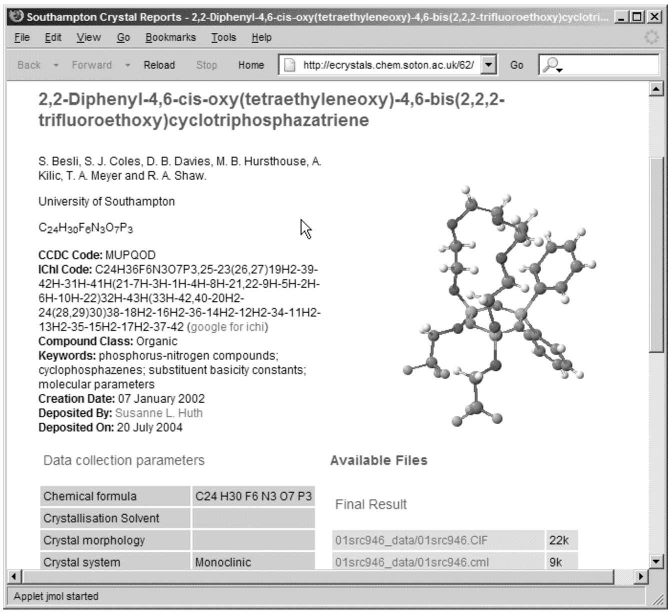
The National Crystallography Service (NCS) in the UK has been constructing a pilot data repository of this type over the last two years as part of the JISC-funded eBank project [8]. The objective is to create a managed repository of data sets, structure models and workbench reports associated with each experiment conducted by the NCS. These resources provide the supporting materials for work that is published or formally deposited in structural databases, but it also provides a home for structural studies that, for one reason or another, do not find their way into a publication. The management team behind this project are exploring ways of ensuring that such information can find its way into the public domain if they are not published within a reasonable length of time. The issues connected with this are not technical, but have to do with intellectual property rights and licensing of data collected on behalf of clients of the NCS.

Initiatives such as this have already won a measure of support within the crystallographic community, where there is much concern about experimental results failing to appear in the public domain. However, there are some worries about the quality of the data available from such repositories, where such data have not undergone the peer review associated with formal publication. Part of the IUCr's involvement with such initiatives has been to promote the generation of standard *checkcif* reports for the collected structures (using the same public service that we developed for journal peer review), and to offer the *checkcif* report as a quality indicator among the supporting metadata available from eBank for each structure.

14. Distributed publishing resources

The eBank project of the NCS provides not merely a rather 'dark' data depository, but rather an interactive metadata-rich graphical interface to a publication system based on the EPrints software engine (http://software.eprints.org/) used by a growing number of institutional repositories (Fig. 7).

In the typical user interface to a structure report as shown in Fig. 7, the underlying structure is characterized from a set of metadata elements summarizing its chemistry, crystal symmetry and other useful attributes. The IUCr has also been working with the eBank team to determine a standard set of metadata elements that could be adopted by similar initiatives to provide a common characterization for use by software indexing and search engines.


Southampton Crystal Reports - 2,2-Diphenyl-4,6-cis-oxy(tetraethyleneoxy)-4,6-bis(2,2,2-trifluoroethoxy)cyclotri...
File
Edit
View
Go
Bookmarks
Tools
Help
Back
Forward
Reload
Stop
Home
http://ecrystals.chem.soton.ac.uk/62/
Go

2,2-Diphenyl-4,6-cis-oxy(tetraethyleneoxy)-4,6-bis(2,2,2-trifluoroethoxy)cyclotriphosphazatriene

S. Besli, S. J. Coles, D. B. Davies, M. B. Hursthouse, A. Kilic, T. A. Meyer and R. A. Shaw.

University of Southampton

$C_{24}H_{30}F_6N_3O_7P_3$

CCDC Code: MUPQOD
IChI Code: C24H36F6N3O7P3,25-23(26,27)19H2-39-42H-31H-41H(21-7H-3H-1H-4H-8H-21,22-9H-5H-2H-6H-10H-22)32H-43H(33H-42,40-20H2-24(28,29)30)38-18H2-16H2-36-14H2-12H2-34-11H2-13H2-35-15H2-17H2-37-42 (google for ichi)
Compound Class: Organic
Keywords: phosphorus-nitrogen compounds; cyclophosphazenes; substituent basicity constants; molecular parameters
Creation Date: 07 January 2002
Deposited By: Susanne L. Huth
Deposited On: 20 July 2004

Data collection parameters

Chemical formula	C24 H30 F6 N3 O7 P3
Crystallisation Solvent	
Crystal morphology	
Crystal system	Monoclinic

Available Files

Final Result

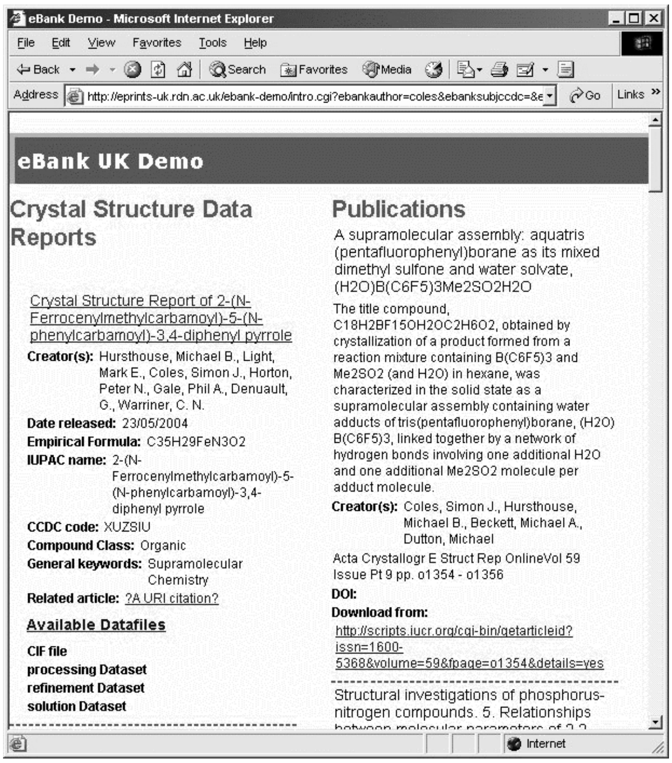
01src946_data/01src946.CIF	22k
01src946_data/01src946.cml	9k

Applet jmol started

Fig. 7: A crystal structure report as presented through the eBank user interface.

An important component of this project is the ability to link effectively between records in the repository and related database records or publications. Fig. 7 demonstrates how structures deposited in the Crystal Structure Database of the Cambridge Crystallographic Data Centre (CCDC) include the CCDC identifier code as part of the standard metadata reported to the user. In a further pilot, the project has demonstrated the ability to link records from the repository directly to publications (Fig. 8). For publications in IUCr journals, it is our intention to develop reciprocal linking from the published article back to the eBank holding.

We see this as the first stage in a distributed publication system for crystallography that goes beyond the traditional peer-reviewed journal. By providing effective links and query mechanisms, it becomes possible to provide the researcher with access not only to published results, but to the original data (including diffraction images that are too large for storage by publishers), to unpublished structures, to database holdings, and in all cases to be able to assess the quality of the reported results.


eBank Demo - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Back
Search
Favorites
Media
Address
http://eprints-uk.rdn.ac.uk/ebank-demo/intro.cgi?ebankauthor=coles&ebanksubjccdc=&e
Go
Links »
Internet

eBank UK Demo

Crystal Structure Data Reports

Crystal Structure Report of 2-(N-Ferrocenylmethylcarbamoyl)-5-(N-phenylcarbamoyl)-3,4-diphenyl pyrrole

Creator(s): Hursthouse, Michael B., Light, Mark E., Coles, Simon J., Horton, Peter N., Gale, Phil A., Denuault, G., Warriner, C. N.

Date released: 23/05/2004

Empirical Formula: C35H29FeN3O2

IUPAC name: 2-(N-Ferrocenylmethylcarbamoyl)-5-(N-phenylcarbamoyl)-3,4-diphenyl pyrrole

CCDC code: XUZSIU

Compound Class: Organic

General keywords: Supramolecular Chemistry

Related article: ?A URI citation?

Available Datafiles

CIF file
processing Dataset
refinement Dataset
solution Dataset

Publications

A supramolecular assembly: aquatris (pentafluorophenyl)borane as its mixed dimethyl sulfone and water solvate, (H2O)B(C6F5)3Me2SO2H2O

The title compound, C18H2BF15OH2OC2H6O2, obtained by crystallization of a product formed from a reaction mixture containing B(C6F5)3 and Me2SO2 (and H2O) in hexane, was characterized in the solid state as a supramolecular assembly containing water adducts of tris(pentafluorophenyl)borane, (H2O) B(C6F5)3, linked together by a network of hydrogen bonds involving one additional H2O and one additional Me2SO2 molecule per adduct molecule.

Creator(s): Coles, Simon J., Hursthouse, Michael B., Beckett, Michael A., Dutton, Michael

Acta Crystallogr E Struct Rep OnlineVol 59 Issue Pt 9 pp. o1354 - o1356

DOI:

Download from: http://scripts.iucr.org/cgi-bin/getarticleid?issn=1600-5368&volume=59&fpage=o1354&details=yes

Structural investigations of phosphorus-nitrogen compounds. 5. Relationships

Fig. 8: A prototype results interface for an eBank query, returning results both in the eBank holding and the published literature.

15. Resource discovery

The first steps towards such a distributed publication system have been taken through active collaboration between eBank and partners such as the IUCr. However, it will certainly be the case that other groups will wish to self-archive data, from the regional scale (*e.g.* the NSF-funded Reciprocal Net project www.reciprocalnet.org in the United States) down to the individual researcher who wishes to manage a personal data collection. Such a diversity of repositories will inevitably be managed by different software engines, and will have different levels of functionality. Two things are needed to tie together such a disparate collection of resources.

One is simply knowledge that such resources exist. At a minimum, they should support a common protocol that allows resource discovery and identification. A candidate protocol for this purpose is OAI-PMH, the Protocol for Metadata Harvesting of the Open Archives Initiative [14]. It has been designed specifically for metadata exposure in an open archives framework, and contains useful hooks for characterizing collections within an archive (therefore allowing selection of say 'crystal data sets' from a university-wide repository of academic content), and for advertising the existence of known related resources.

OAI-PMH is already an integral part of the EPrints software that underpins eBank, and is relatively easy to implement either within existing software or as a standalone interface layer.

An important component of the protocol is its meticulous timestamping of metadata, so that it can be used to track changes in an advertised collection. We have already noted in Section 6 above its potential for synchronizing current content with a growing archive, using precisely this feature. We can therefore envisage that a network of data repositories linked by OAI-PMH could be harvested by a central aggregator licensed to make an archive copy of all the material placed in the public domain through the individual repositories. In principle, such a centralized archive could safeguard access in perpetuity to the large and growing number of data sets that are of high quality, properly evaluated, and yet never formally published.

16. Distributed querying

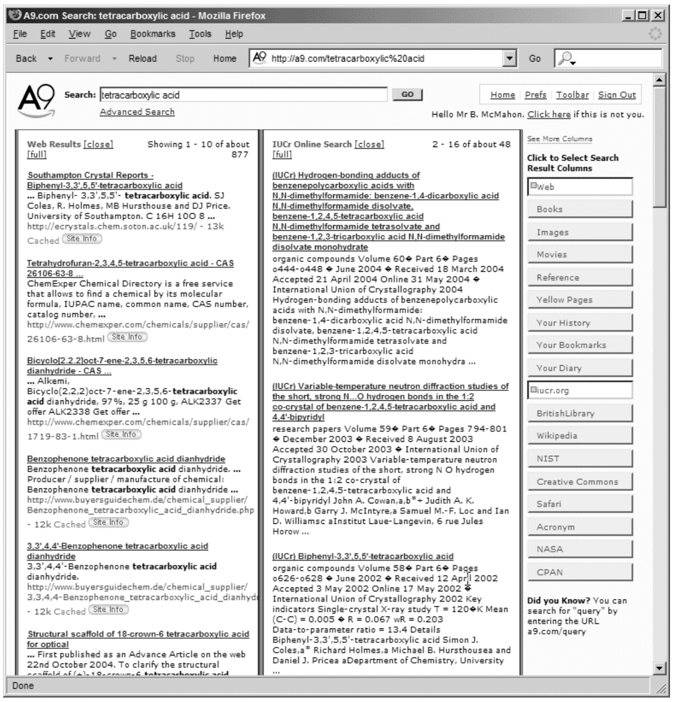
A centralized aggregator or archive could also provide a centralized search engine for crystallographic data. At first sight this offers a potential threat to existing structural databases such as the CCDC. In practice, however, the existing databases offer a range of very sophisticated query facilities based on scientific analysis and on highly optimized data representations, and it is by no means clear that a generalized aggregator could compete on these terms.

Even without aggregation, however, it would be attractive to be able to send a crystallographic query to all the available data sources and receive a composite response. Distributed querying of full-text collections is becoming a real possibility, for example using the OpenSearch protocol developed by Amazon for its A9 search engine [1]. Fig. 9 demonstrates a search on a chemical name that returns both the eBank record as indexed by Google and a link to the associated article as queried directly on the IUCr journals site.

This is a powerful facility, but would be even more powerful if structured queries could be made on numeric quantities within a data set. There are pressures to extend the OpenSearch protocol by adding common metadata fields to which searches could be restricted. In principle, such metadata elements could be extended to any agreed set of scientific metadata terms.

As a parenthesis, it should be recognized that in certain cases implicit metadata-based searches can already be carried out on Google or other public full-text search engines. For example, Coles *et al.* [7] have demonstrated that one may simply enter an InChI string in a Google search to locate documents containing the InChI identifier. (InChI is the International Chemical Identifier sponsored by the International Union of Pure and Applied Chemistry [16].) The InChI is a standard metadata item stored in the eBank repository, and Fig. 7 demonstrates how the eBank user interface helpfully provides a direct link to a Google search on the InChI associated with the current structure.

In a similar way, a knowledgeable crystallographer could take account of the particular structure of data-name tags in the CIF format to perform a Google search on the string '_cell_length_b 8.0685'. The result would include any indexed data sets in CIF format reporting a crystal unit cell of length 8.0685 ångströms. (This example has been deliberately chosen to return, among a small number of other hits, the CIF data from eBank for the example structure of Fig. 7.)


A9.com Search: tetracarboxylic acid - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Back Forward Reload Stop Home
Go
A9
GO
Home | Prefs | Toolbar | Sign Out
See More Columns
Click to Select Search Result Columns
□Web
Books
Images
Movies
Reference
Yellow Pages
Your History
Your Bookmarks
Your Diary
□iucr.org
BritishLibrary
Wikipedia
NIST
Creative Commons
Safari
Acronym
NASA
CPAN
Done

http://a9.com/tetracarboxylic%20acid

Search: tetracarboxylic acid

Advanced Search

Hello Mr B. McMahon. Click here if this is not you.

Web Results [close] [full] Showing 1 - 10 of about 877

Southampton Crystal Reports - Biphenyl-3,3',5,5'-tetracarboxylic acid
... Biphenyl- 3,3',5,5'- **tetracarboxylic acid.** SJ Coles, R. Holmes, MB Hursthouse and DJ Price. University of Southampton. C 16H 10O 8 ...
http://ecrystals.chem.soton.ac.uk/119/ - 13k Cached Site Info

Tetrahydrofuran-2,3,4,5-tetracarboxylic acid - CAS 26106-63-8 ...
ChemExper Chemical Directory is a free service that allows to find a chemical by its molecular formula, IUPAC name, common name, CAS number, catalog number, ...
http://www.chemexper.com/chemicals/supplier/cas/26106-63-8.html Site Info

Bicyclo[2.2.2]oct-7-ene-2,3,5,6-tetracarboxylic dianhydride - CAS ...
... Alkemi, Bicyclo(2,2,2)oct-7-ene-2,3,5,6-**tetracarboxylic acid** dianhydride, 97%, 25 g 100 g, ALK2337 Get offer ALK2338 Get offer ...
http://www.chemexper.com/chemicals/supplier/cas/1719-83-1.html Site Info

Benzophenone tetracarboxylic acid dianhydride
Benzophenone **tetracarboxylic acid** dianhydride. ... Producer / supplier / manufacture of chemical: Benzophenone **tetracarboxylic acid** dianhydride ...
http://www.buyersguidechem.de/chemical_supplier/Benzophenone_tetracarboxylic_acid_dianhydride.php - 12k Cached Site Info

3,3',4,4'-Benzophenone tetracarboxylic acid dianhydride
3,3',4,4'-Benzophenone **tetracarboxylic acid** dianhydride.
http://www.buyersguidechem.de/chemical_supplier/3,3,4,4-Benzophenone_tetracarboxylic_acid_dianhyd - 12k Cached Site Info

Structural scaffold of 18-crown-6 tetracarboxylic acid for optical
... First published as an Advance Article on the web 22nd October 2004. To clarify the structural scaffold of (±)-18-crown-6 **tetracarboxylic acid**

IUCr Online Search [close] [full] 2 - 16 of about 48

(IUCr) Hydrogen-bonding adducts of benzenepolycarboxylic acids with N,N-dimethylformamide: benzene-1,4-dicarboxylic acid N,N-dimethylformamide disolvate, benzene-1,2,4,5-tetracarboxylic acid N,N-dimethylformamide tetrasolvate and benzene-1,2,3-tricarboxylic acid N,N-dimethylformamide disolvate monohydrate
organic compounds Volume 60� Part 6� Pages o444-o448 � June 2004 � Received 18 March 2004 Accepted 21 April 2004 Online 31 May 2004 � International Union of Crystallography 2004 Hydrogen-bonding adducts of benzenepolycarboxylic acids with N,N-dimethylformamide: benzene-1,4-dicarboxylic acid N,N-dimethylformamide disolvate, benzene-1,2,4,5-tetracarboxylic acid N,N-dimethylformamide tetrasolvate and benzene-1,2,3-tricarboxylic acid N,N-dimethylformamide disolvate monohydra ...

(IUCr) Variable-temperature neutron diffraction studies of the short, strong N...O hydrogen bonds in the 1:2 co-crystal of benzene-1,2,4,5-tetracarboxylic acid and 4,4'-bipyridyl
research papers Volume 59� Part 6� Pages 794-801 � December 2003 � Received 8 August 2003 Accepted 30 October 2003 � International Union of Crystallography 2003 Variable-temperature neutron diffraction studies of the short, strong N O hydrogen bonds in the 1:2 co-crystal of benzene-1,2,4,5-tetracarboxylic acid and 4,4'-bipyridyl John A. Cowan,a,b*+ Judith A. K. Howard,b Garry J. McIntyre,a Samuel M.-F. Loc and Ian D. Williamsc aInstitut Laue-Langevin, 6 rue Jules Horow ...

(IUCr) Biphenyl-3,3',5,5'-tetracarboxylic acid
organic compounds Volume 58� Part 6� Pages o626-o628 � June 2002 � Received 12 April 2002 Accepted 3 May 2002 Online 17 May 2002 � International Union of Crystallography 2002 Key indicators Single-crystal X-ray study T = 120�K Mean (C-C) = 0.005 � R = 0.067 wR = 0.203 Data-to-parameter ratio = 13.4 Details Biphenyl-3,3',5,5'-tetracarboxylic acid Simon J. Coles,a* Richard Holmes,a Michael B. Hursthousea and Daniel J. Pricea aDepartment of Chemistry, University ...

Did you Know? You can search for "query" by entering the URL a9.com/query

Fig. 9: Syndicated query response from A9 for the term 'tetracarboxylic acid'. A Southampton eBank record is located by Google (top of the left-hand column); the associated article is also found in a simultaneous search of the IUCr journals site (bottom of right-hand column).

17. Necessary steps towards realization

We have indicated how large amounts of scientific and other metadata are generated and transferred within the CIFs that document a crystallographic experiment. We have demonstrated that these can be easily extracted to populate bibliographic databases and scientific databanks. We have also given some indication of how effective such metadata can be in driving workflow, whether at the editorial office or the laboratory workbench. We have presented a vision of distributed scientific information resources linking publication, database and laboratory archive, all tied together by appropriate metadata.

To realize the vision, two steps are necessary. One is the orderly assignment of identifiers so that disparate resources (publications, structural models, images) can be effectively linked together. The other is the agreement of standard metadata populated by all the participants in the maintenance of such distributed resources.

The natural candidate for an identifier is the DOI. We have described our efforts within CrossRef to assign DOIs to individual supplementary data sets. However, CrossRef is currently a publishers' organization, and may not be seen as a natural registration agency for scientific data. Registration agencies already exist who do actively register DOIs for scientific data sets. For example, the Technische Informationsbibliothek Hannover (www.tib.uni-hannover.de) assigns DOIs to data sets in meteorology and earth sciences. However, if multiple registration agencies are registering scientific data, it is clear that they will need common metadata to characterize the individual data sets and so ensure interoperability.

Thus the promotion of scientific metadata standards is a key step in achieving the vision we have outlined above, not only in the field of crystallography, but potentially across all the sciences. What is required is a common set of metadata characterizing the nature and properties of scientific data sets, at a level intermediate between finely-grained discipline-specific collections (such as the CIF dictionaries) and the rather coarse-grained standards, such as Dublin Core, currently in use in the publishing field. The Dublin Core Metadata Initiative specification [10] is, in many ways, a useful starting point, both because it has been very successful in the publishing world despite its limitations, and because it draws attention specifically to the topics of provenance and rights that will be essential in permitting distributed management of scientific data from different sources, and which we have not discussed here in any detail.

In addition to metadata concerning provenance and rights, a scientific metadata scheme should be capable of identifying the field of science to which a data set relates, whether it is static or time-series data, what other data records it is related to, and probably a broad indication of the type of experiment. Compiling a comprehensive scheme that will win acceptance across the wide spectrum of the sciences will not be an easy task, and it may properly involve the active participation of Scientific Unions, inter-Union organizations such as

CODATA (http://www.codata.org), and scientific standards organizations.

It will be a lengthy journey to achieve the objective of a common scientific metadata specification and a network of distributed resources interacting through protocols built on such metadata. However, our experiences as an electronic publisher in the last few years and our collaborations with the data community demonstrate that within the field of crystallography the journey has definitely begun.

References

[1] A9.com, Inc. 'OpenSearch.' http://opensearch.a9.com/ (2005).

[2] Allen, F. H. 'The Cambridge Structural Database: a quarter of a million crystal structures and rising', *Acta Crystallogr. Section B*, 58, pp. 380-388, (2002).

[3] Anklesaria, F., McCahill, M., Lindner, P., Johnson, D., Torrey, D., Alberti, B. 'The Internet gopher protocol (a distributed document search and retrieval protocol)', RFC 1436. Network Working Group. http://www.ietf.org/rfc/rfc1436.txt, (1993).

[4] ANSI/NISO. 'Information retrieval (Z39.50): Application service definition and protocol specification. Z39.50-1995', ttp://lcweb.loc.gov/z3950/agency/1995doce.html, (1995).

[5] ANSI/NISO. 'Serial Item and Contribution Identifier (SICI) Z39.56-1996 (R2002)', (1996).

[6] Bourne, P. E., Berman, H. M., McMahon, B., Watenpaugh, K. D., Westbrook, J. D., Fitzgerald, P. M. D. 'Macromolecular Crystallographic Information File', *Methods Enzymol.* 277, pp. 571-590, (1997).

[7] Coles, S., Lyon, L., Carr, L., Heery, R., Hursthouse, M., Gutteridge, C., Duke, M., Frey, J., Roure, D. 'eBank UK Linking Research Data, Scholarly Communication and Learning', in *Semantic Grid Workshop*, Global Grid Forum 11, Hawaii, USA, 4-7 July 2004. Hawaii, Global Grid Forum, 8pp (in press). Preprint at http://eprints.soton.ac.uk/12461/, (2004)

[8] Coles, S. J., Day, N. E., Murray-Rust, P., Rzepa, H. S., Zhang, Y. 'Enhancement of the chemical semantic web through the use of InChI identifiers', *Org. Biomol. Chem.*, 3, pp. 1832-1834, (2005).

[9] CrossRef 'CrossRef XML Schema: deposit of journal, book, & conference proceedings metadata, Version 3.0.3', http://www.crossref.org/depositSchema/crossref3.0.3.xsd, (2004).

[10] DCMI Usage Board, 'DCMI Metadata Terms', http://dublincore.org/documents/dcmi-terms/, (2005).

[11] Hall, S. R., Allen, F. H., Brown, I. D. 'The Crystallographic Information File (CIF): a new standard archive file for crystallography', *Acta Crystallogr. Section A*, 47, pp. 655-685. (1991).

[12] *International Tables for Crystallography* Vol. G, 'Definition and Exchange of Crystallographic Data', edited by S. R. Hall & B. McMahon. Berlin: Springer, (2005).

[13] Kahle, B. 'An information system for corporate users: wide area information servers', Thinking Machines technical report TMC-199. See also *Online*, 15, pp. 56-60, (1991).

[14] Open Archives Initiative, 'The Open Archives Initiative Protocol for Metadata Harvesting, version 2.0 of 2002-06-14', http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm, (2002).

[15] Research Collaboratory for Structural Bioinformatics 'PDB Annual Report 2004. http://www.rcsb.org/pdb/annual_report04.pdf, (2004).

[16] Stein, S. E., Heller, S. R., Tchekhovskoi, D. 'An Open Standard for Chemical Structure Representation: The IUPAC Chemical Identifier', in 'Proceedings of the 2003 International Chemical Information Conference (Nimes)', *Infonortics*, pp. 131-143, (2003).