Journal of
Applied Crystallography
ISSN 0021-8898

Received 13 September 2011
Accepted 13 February 2012

CrystalEye: automated aggregation, semantification and dissemination of the world's open crystallographic data

Nick Day, Jim Downing, Sam Adams, N. W. England and Peter Murray-Rust*

Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK. Correspondence e-mail: pm286@cam.ac.uk

CrystalEye automatically aggregates crystallographic data from web resources (the supplementary data to articles on publishers' web sites) into a structured XML-based repository, and then adds value to that open data by providing methods to easily browse, search and to keep up-to-date with the latest published information.

1. Introduction

Crystallography has led the world in the systematic capture, validation and dissemination of the results of experiments, particularly structure determination. The impetus came largely from J. D. Bernal who had the vision of a crystallographic factory to capture and disseminate structural information on a global scale (Bernal, 1965). This vision was taken up by the Cambridge Crystallographic Data Centre (CCDC), led by Olga Kennard, culminating in the creation of the Cambridge Structural Database (CSD), which is seen as one of the pioneering scientific databases (Allen *et al.*, 1979). Initially the data were collected by hand (abstracted by humans from the pages of printed journals), but more recently electronic input in the form of CIFs has become the norm. The CIF, or Crystallographic Information File, is a structured document format with a formal syntax and associated ontologies, designed for the exchange of crystallographic information (Hall *et al.*, 1991). The International Union of Crystallography (IUCr) has campaigned vigorously to support the view that all crystallographic structure determinations should report the associated data in CIF form, and compliance is now very high. In some cases authors are required to send their CIFs to the journal directly, which then will make them available on web pages after publication, and in other cases the journals require the data to be transmitted to the CCDC for deposition.

The increasing quality of CIF data files and the high compliance means that effectively all published small-molecule crystallographic data are available publicly in CIF form. Until recently these data had to be managed through the CCDC deposition process, but the ubiquity of science publishing in electronic journals with associated data files means that it is technically possible to extract this information by modern web technologies, and *CrystalEye* is based on this process.

The rapidly increasing amount of information ('the data deluge') (Hey *et al.*, 2009) has led to major socio-political discussions about the provision of, ownership of and intellectual property contained in scientific data. Traditionally it has been reasonable that human abstraction from the scientific literature should be rewarded by requiring readers to pay for monographs or access to databases [Landholt-Bernstein (http://www.springermaterials.com/navigation/) *etc.*]. However, the power of modern technology and web tools means that, in many cases, crystallography being a pre-eminent example, it is possible to devise automated systems that run without significant human input. This changes the balance between author, data abstractor, database provider and reader. *CrystalEye* is therefore not only a valuable crystallographic resource in itself but a wider demonstration of our ability to aggregate the world's scientific knowledge in new and more flexible ways.

The major socio-political problem at the moment is the availability of crystallographic data associated with papers from certain publishers. Using Stallman's terminology, these data are free (as in beer) but not free (as in speech) (http://www.gnu.org/philosophy/free-sw.html). It is possible to get small amounts of information on specific request to the CCDC, but it is not clear what the redistribution rights are, and requests for large amounts of information are unlikely to be successful. *CrystalEye*, therefore, demonstrates the difference between completely open data (as in the Open Knowledge Foundation Open Definition; http://www.opendefinition.org/) and data that are only available under specific terms, which may be subject to change in the future. One of us (PM-R) has written to scientific publishers and journal editors requesting that their CIFs be made publicly available as part of the primary scientific record. Some major scientific publishers [IUCr, Nature, Royal Society of Chemistry (RSC) and American Chemical Society (ACS)] do make CIFs publicly available, but those who do not completely and publicly release CIF data include Elsevier, Wiley and Springer.

The *CrystalEye* vision is based upon the principle that copyright protection is only afforded to materials that represent an author's creative work, not to raw scientific data. The IUCr specifically states (http://journals.iucr.org/services/authorrights.html) 'Copyright protection is not extended to files of scientific data (*e.g.* structural data CIFs, structure factors, primary diffraction images), and such data sets may be used for *bona fide* research purposes within the scientific community so long as proper attribution is given to the source from which they were obtained.'

Since the raw data are not protected by copyright, they may be re-used by *CrystalEye* without restriction.

1.1. The history of *CrystalEye*

Innovation based on data is difficult unless the innovator has access to all the data and the freedom to do whatever they wish with it and to redistribute it in whatever form they wish. This is not possible by default with data from the CCDC. The *CrystalEye* system described here therefore represents a large resource for innovation in a world that is increasingly using linked open data as a way of unifying and diversifying scientific knowledge.

CrystalEye was first created as a by-product of a thesis (Day, 2009). The primary purpose of the thesis was to investigate whether semi-empirical calculations (*e.g. MOPAC*; http://openmopac.net/) could reproduce some or all of the observed geometry of crystal structures. Because *MOPAC* has been tuned against many types of structure, it was important to have a large number of high-quality structures to test against. We therefore decided to make a complete collection of organic, organometallic and inorganic structures in a single source.

Although there are some open collections [such as the Crystallography Open Database (COD; Gražulis *et al.*, 2009)], we felt that the most cost-effective and appropriate means would be to collect structures from the then-current literature so that we could associate the data with a publication. Therefore, we developed automatic 'crawling' tools (pubcrawler) that could download and aggregate large amounts of current material. Over time, we have also extracted earlier papers so that *CrystalEye* is now a record of a substantial portion of the publicly available crystallographic data. We note in passing that the quality of structures has improved over the period 1991–2011, both in their reported quality criteria and in the compliance of the CIFs (content and syntax).

Because we were repeatedly comparing structures with calculations, we needed search tools for our collection and the ability to identify interesting or outlying structures, and so evolved a searchable repository. The original intention was not to create a formal repository and the data were stored as flat files on disk for convenience. This has proved to be invaluable for the current system (*CrystalEye*, accessible at http://wwmm.ch.cam.ac.uk/crystaleye/), but recently we have needed to re-factor the approach (*CrystalEye2*, accessible at http://crystaleye.ch.cam.ac.uk/). This new system was presented at the 22nd IUCr Congress in Madrid (22–30 August 2011; Microsymposium 89 'Archiving, Exchange and Retrieval of Scientific Data in the 21st Century').

2. *CrystalEye*

CrystalEye automatically identifies, aggregates and indexes openly available CIFs, primarily from the published chemical literature, with almost 100% recall. Additional information including chemical identifiers, structure diagrams, interactive models and Chemical Markup Language (CML) representations of the data are generated, alongside aggregate reports. The original data files and the generated resources are made freely available to browse, search, download and re-use using the latest web standards and technologies, making it much simpler for a researcher to keep updated with the latest relevant published crystallography.

2.1. The *CrystalEye* workflow

The sole focus of *CrystalEye* is on openly available crystallographic data. The majority of *CrystalEye*'s data are currently sourced from the published chemical literature. A number of publishers make CIFs freely available as supplementary data files, and *CrystalEye* automatically locates these using a component known as pubcrawler. Pubcrawler is a bespoke web crawler (or 'spider') that visits the current issue table of contents (TOC) pages of a number of crystallography containing journals published by the ACS, the Chemical Society of Japan, the IUCr, Nature and the RSC. A full list of the journals covered is given in Appendix *A*, available as supplementary material.[1]

[1] Supplementary material is available from the IUCr electronic archives (Reference: HE5533). Services for accessing this material are described at the back of the journal.

Pubcrawler contains custom logic for each publisher's site to identify and collect any CIFs (Fig. 1). While this approach is not easily scalable for generic web crawling, in the limited domain of crystallographic data publishers it is quite feasible, and indeed necessary since the structure of journals' web sites and the location of the supporting information vary between publishers. For example, the IUCr journals provide a link directly from the TOC web page (as shown in Fig. 2 below), whilst RSC journals embed the supporting information URL within the abstract of the specific article. Once located, the CIFs are downloaded and the digital object identifier (DOI) of their source article recorded, ready for the next stage of the process.

It is worth noting that, while this process is automated, some degree of monitoring and maintenance is required – any time a journal changes the structure of its web site it is necessary to update, or in extreme cases, re-write, the associated web spider. For example, the RSC access to supplemental pages is not a simple hyperlink but a complex JavaScript, which we have not parsed, and as a result RSC structures are no longer automatically included incrementally.

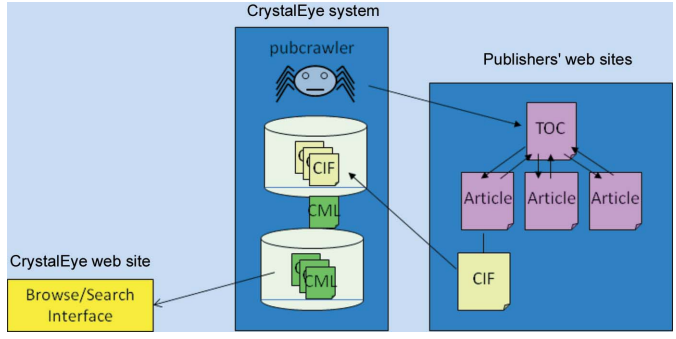

Figure 1
Workflow showing the browsing, aggregation and semantification of *CrystalEye* content. The spider (pubcrawler) visits publishers' web sites, finds the TOC from the latest issue of each journal, extracts each article, determines whether any have supplemental data (CIF) and converts them into CML for browsing/searching *via* the *CrystalEye* web site (http://wwmm.ch.cam.ac.uk/crystaleye/).

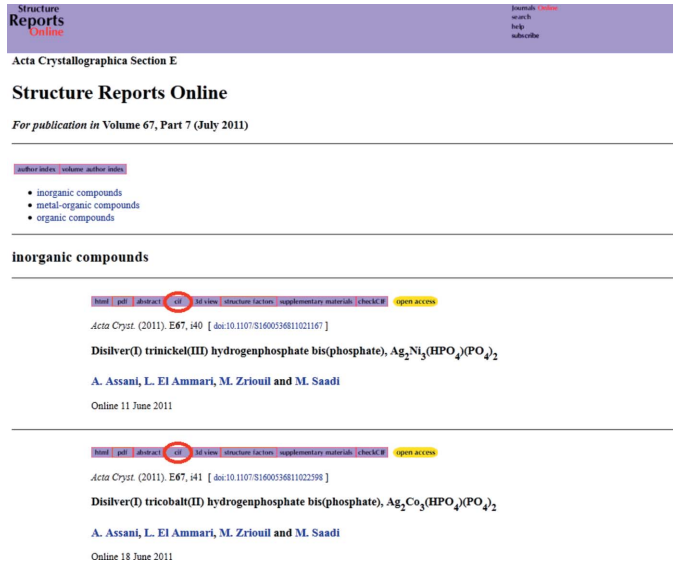

Figure 2
A typical TOC for an issue of *Acta Crystallographica Section E*, which *CrystalEye* browses for individual articles. In this case the crystal data (CIF) is consistently identified through syntax and a customized link (button), highlighted.

2.2. Semantification – *CIFXOM* and JUMBO converters

Once pubcrawler has identified and downloaded newly published CIFs they undergo a process of semantification, transforming their content into a form that is understandable by machines. This process is fully automated and highly reliable. The error rate in the conversion of valid CIFs to CML is less than 0.2%, and the proportion of published open CIFs that are valid with respect to the IUCr specification was around 99% in 2007 and is shown to be improving.

In the initial stage of the semantification process (Fig. 3), the downloaded CIFs are read using *CIFXOM* and converted into the XML-based CIFXML representation (Day *et al.*, 2011). *CIFXOM* is based on the SAX parser/handler model, and is able to correct many common formatting errors and convert the potentially malformed CIFs into syntactically guaranteed XML. *CIFXOM* generates XML with CIF-based semantics (*i.e.* linking back to the IUCr and other local dictionaries) but does not provide enough explicit semantics for re-use in a chemical context (*i.e. via* CML).

A CIF may contain data and metadata referring to more than one crystal structure. In this case, the individual data blocks are identified in the CIFXML data structure, using standard XML data processing tools (XPath), and written into separate files together with a copy of the global data block, if present.

Subsequently, a two-stage process using JUMBOConverters (https://bitbucket.org/wwmm/jumbo-converters) losslessly transforms each CIFXML representation into a complete semantic CML representation with references to standard dictionaries and standardized layout. In the first stage the CIFXML file is converted into 'raw' (syntactic) CML. In this process the structure of the CIF data model is conserved but represented using the CML vocabulary. Properties are referenced against machine-understandable CML dictionaries, data types and units are added, and error values recorded. Next the 'raw' CML is converted into 'complete' (semantic) CML; the data model is restructured to reflect CML conventions and additional 'chemical' information added, allowing interoperability with the large number of CML-supporting tools.

During the conversion from 'raw' to 'complete' CML the connection tables of the chemical structures are derived from the purely atomistic representations published in the CIFs. The machine generation of chemical connection tables from CIFs takes up a large part of the CIF-to-CML conversion. The process requires many heuristics to resolve disorder, identify connected atoms, and assign bond orders and formal charges to atoms, and in some cases it is not possible to deduce the exact connection table. Once the connection table has been generated, disconnected components in the structure are identified and marked as separate moieties in the CML. Most downstream processing will benefit from 'chemical' information added to the 'raw' CML. This includes concepts such as bonds and bond orders, and derived information such as SMILES (http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html) and InChI (International Chemical Identifier; http://www.iupac.org/inchi/) representations of chemical composition and connectivity.

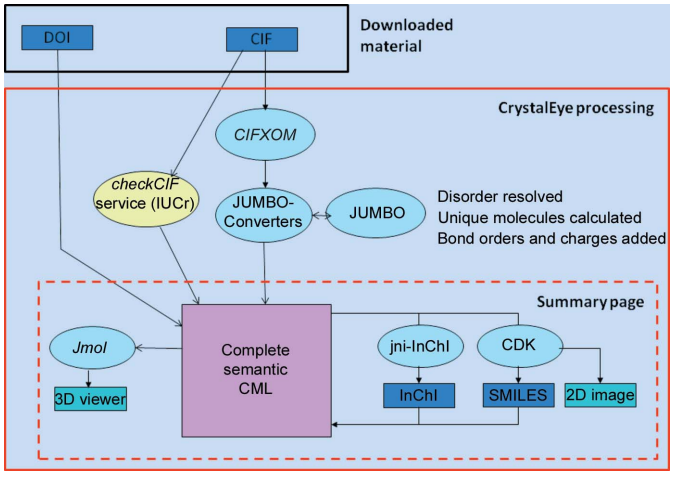

Figure 3
The *CrystalEye* validation and semantification workflow. The downloaded CIF is converted to CIFXML and then, in a two-step process *via* JUMBOConverters, to fully semantic CML. In addition, chemical indexes and two/three-dimensional structures are generated and displayed on the summary page.

The conversion from 'raw' to 'complete' CML consists of the following steps:

(*a*) Interpretation of disorder. Where only occupancies are given, this is heuristic; more recent CIFs often have well defined disorder groups. Where disorder cannot be resolved, further semantics are not always added.

(*b*) Expansion of content to one unit cell with additional neighbours to guarantee identification of all contacts.

(*c*) Calculation of bonded atoms. Using covalent radii works well for a large number of 'molecular' complexes. Beyond that, 'polymeric' complexes along one, two or all three axes are common. In these cases chemical heuristics are highly subjective.

(*d*) Identification of disjoint chemical moieties. These moieties are arranged as molecule children of a single parent CML molecule.

(*e*) Calculation of bond orders. This is subjective and involves complex heuristics utilizing our perception of normal connectivity, aromaticity and metal charge. Some of the interpretation is aided by de-bonding metals (a similar approach is taken during InChI generation).

(*f*) Calculation of two-dimensional coordinates for display. We use libraries such as the Chemistry Development Kit (CDK; Steinbeck *et al.*, 2003) to generate two-dimensional coordinates for every moiety. For all except complex molecules such as buckyballs, the plot is usually informative and attractive.

(*g*) Calculation of formal atom parity stereochemical indicators for C atoms. Note that atom parities are an objective indicator similar to those found in SMILES and InChI. We can also, if necessary, apply limited Cahn–Ingold–Prelog rules to generate *R*- and *S*- labels.

(*h*) Calculation of bond stereochemical indicators (*cis*/*trans* and *E*/*Z*). Again bond stereochemistry is an objective indicator: *E*/*Z* and *cis*/*trans* depend on stereochemical rules.

(*i*) Calculation of compositional formula by summing atom occupancies.

(*j*) Calculation of SMILES formula (not canonical).

At the end of this the chemistry is fit for input into CML-compliant software and general chemical indexing and searching.

The overall semantification process can be summarized as follows:

(*a*) Creation of raw CML. This involves the objective algorithmic conversion of CIFXML to CML without subjective chemical judgement.

(*b*) Conversion of data types. CIF provides only basic data types (char and numb) – these are heuristically enhanced to provide CML/World Wide Web Consortium data types (string, integer, double, boolean, date).

(*c*) Conversion of data structures. CIF provides only item (scalar) and loop (one-dimensional array) structures. There are many implicit semantics (*e.g.* subscripted array and matrix elements). These are heuristically combined into CML's scalar, array and matrix elements (which allow for syntactic definition and checking of sizes, data types, delimiters *etc.*).

(*d*) Conversion to CML objects. CML provides many higher-level objects such as molecules and cells. This stage of the semantification process includes

(i) conversion of loops with `atom_site` (*etc.*) into CML molecules;

(ii) merging of `atom_site` and `atom_site_aniso` into a single molecule with child properties of atoms;

(iii) adding disorder information as atom children;

(iv) conversion of `cell_length` and `cell_angle` to a single CML crystal object;

(v) conversion of fractional coordinates to Cartesian coordinates.

2.3. Additional resources

Once the conversion from CIF into semantic CML is completed, a number of additional resources are generated to add to the value of the crystallographic data. The CIFs are submitted to the IUCr's *checkCIF* service (http://checkcif.iucr.org/), and the *checkCIF* report and *PLATON*-generated (Spek, 1990) ellipsoid plot archived. For organic structures (but not inorganic or polymeric ones) InChI and InChIKey identifiers are generated using JNI-InChI (http://jni-inchi.sourceforge.net/), SMILES using OpenBabel (http://openbabel.org/), and two-dimensional structure diagrams using the CDK for each overall structure and each moiety.

The *checkCIF* reports and SMILES, InChI and InChIKey identifiers are merged into the 'complete' CML.

2.4. Quality and errors

CrystalEye will attempt to detect errors but not to correct them. It will also indicate where structures are likely to be of high or low quality. For example, if the repository is searched on the bond-length parameter (see §3.2.3), high-quality structures ('after protocol'; Townsend, 2007) are selected if both the temperature and the *R* factor are sufficiently low. Where a structure is disordered, *CrystalEye* will attempt to interpret the basic chemical structure by taking the major component and generating a connection table. All disordered atoms, whether identified as such by the author or deduced from fractional occupancies, are retained in the complete CML, but this may be annotated to indicate the major components.

There is no independent check of chemical constitution as most authors neither report the crystallographic connectivity nor give a systematic chemical name. Recently (unpublished work) we have analysed several thousand structures from *Acta Crystallographic Section E* where there are both systematic names and structural diagrams (provided by the authors) included in the abstract and full text. We have detected almost no cases where the crystal structure coordinates do not correspond to the name and/or the diagram. We cannot be sure that other publishers ensure the same high standard of atomic coordinates owing to their less permissive licences restricting such analyses.

We note that some of this checking would be greatly enhanced if the CIF dictionary `_chemical_conn` data items and systematic name were required when publishing CIFs.

2.5. Statistics of *CrystalEye* contents

CrystalEye was initiated in 2007 and has been running constantly since then. The following ingest statistics refer to 2007:

(*a*) International Union of Crystallography: 30 465.

(*b*) Royal Society of Chemistry: 10 351

(*c*) American Chemical Society: 26 548

(*d*) Elsevier: 168

(*e*) Chemical Society of Japan: 150

(*f*) Crystallography Open Database: 18 283

Since then, we have withdrawn the Elsevier material as it was behind a paywall, and have continued to ingest from the other sources. The system had processed about 150 000 compound CIFs by 2011, which contain about 250 000 individual structures.

The statistics up to 2007 are shown in Fig. 4.

The bond-search indexes were frozen in 2008, as they contained millions of bonds. The generated bond histograms and substructure search (see §3.2.3) give users useful generic results although some of the later structures may not be individually indexed. Similarly the generation of fragments was discontinued because of their very large number and the lack of a suitable index. In our re-factored design (*CrystalEye2*) the generation of search indexes and fragments is formally separated from the creation of the semantic CML.

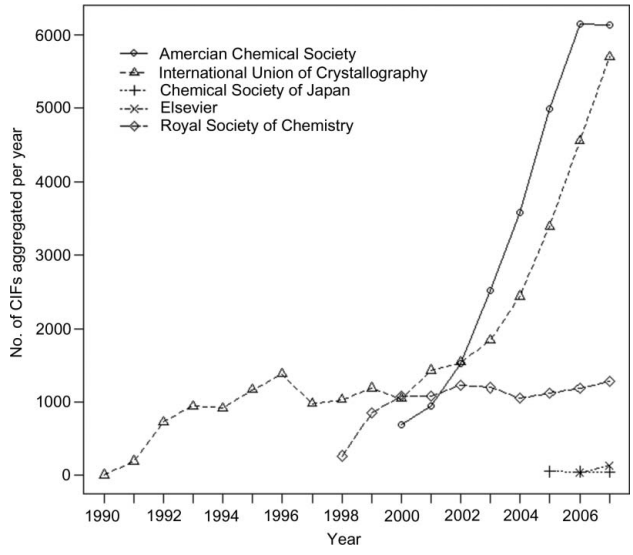

Figure 4
Plot showing the number of CIFs aggregated from each publisher from 1990 to 2007.

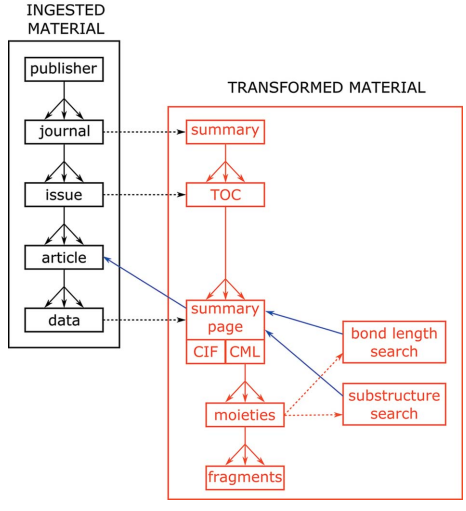

Figure 5
CrystalEye content. The external web pages (left, black) are ingested and translated to journal summaries, issue TOCs, and individual CIFs (no material behind paywalls is ingested), CML and summary pages. The CML is processed hierarchically to moieties and collections of fragments. The searches link to summary pages and the summary pages link to online articles. Solid arrows in the figure represent hyperlinks, dashed arrows processing steps; forked arrows are 1:*n* containers.

3. *CrystalEye* web site and dissemination tools

The structure of the *CrystalEye* web site is shown schematically in Fig. 5.

3.1. Summary web pages

'Summary' web pages are generated for each crystal structure, containing key data, bibliographic information and a link to the original article, and links to the original CIF and the generated files. Both two-dimensional and three-dimensional representations of the structures are provided [using CDK and *Jmol* (http://jmol.sourceforge.net/), respectively].

3.2. Derived data

The summary web pages represent the core of *CrystalEye*. In addition, we have enhanced this information with derived data and indexes.

3.2.1. Moieties. We define a moiety as a discrete sub-graph in the connection table, and it is most useful for organic and organometallic structures. JUMBO is able to determine moieties automatically and the system creates individual pages for each moiety (other than those with single atoms or XH_n molecules, such as water). This leads to millions of pages with fully interactive molecules, which can be viewed in *Jmol* and queried for geometry. There is a linking feature which allows tables of bonds in a moiety to be interrogated either from the structure or from the table. Moieties can be visited in the browser as 'child pages' of the summary web pages.

3.2.2. Fragments. We have also derived fragments within each structure according to the following rules. A moiety without metal atoms is broken down into ring nuclei (*i.e.* substructures where all the bonds are cyclic) which are connected by chains. A moiety containing metal atoms is broken down to metal centres and their ligands. The metal centres are of two types: (1) single atoms and (2) metal clusters.

3.2.3. Search indexes. *CrystalEye* provides a number of ways to search for crystal structures: by cell dimension, by chemical structure and by bond type/length. Crystal structures are indexed at the time of ingest.

(*a*) Cell dimensions. The cell dimensions are subjected to a Delaunay reduction (Patterson & Love, 1957), which allows queries on unit-cell dimensions within a given tolerance. Users may also search on cell volume (again with tolerance).

(*b*) Chemical substructures. The OpenBabel library is used to provide *CrystalEye*'s substructure search functionality. Entries are indexed and searchable by using standard SMILES patterns as input.

(*c*) Bonds. All bonded atom pairs are extracted and indexed by atom types and by their separation (bond length). The bonds for a given pair of types are plotted as an interactive 'clickable' histogram and every entry containing a bond can be reached from the histogram. Thus, a bin of, say, 1.99–2.00 Å will create a summary table of all the structures containing a bond of that length, and these can be browsed in the normal fashion. In addition, the actual bond is highlighted in *Jmol*. There are two indexes, one containing all bonds of a given type and the other only those bonds from 'accurate structures' using the aforementioned protocol.

These concepts are further explained in our screenshot-supported 'tours' of the *CrystalEye* system (see §3.4 and supplementary material).

3.3. Alerting service

Web feeds are widely used to provide users with notifications of updated content, *e.g.* on blogs and news sites. Typically a feed document lists recent content, and by monitoring ('subscribing to') a feed, users can be alerted when new content is published. Entries in a feed document typically contain the title and summary of an item, along with a link to the full resource. An Atom (Atom Publishing Protocol; http://tools.ietf.org/html/rfc5023) feed document can also contain a link to a previous page containing entries describing other content, allowing a client application to discover all the content in the system, not just the most recently updated. *CrystalEye* generates Atom archive feeds, permitting subscribers to receive automatic notifications when new structures are added to the repository.

The following feeds are generated:

(i) All data.

(ii) By journal.

(iii) By compound class: *e.g.* organic, inorganic, organometallic.

(iv) Structures containing each element.

(v) Structures containing a bond between two particular elements.

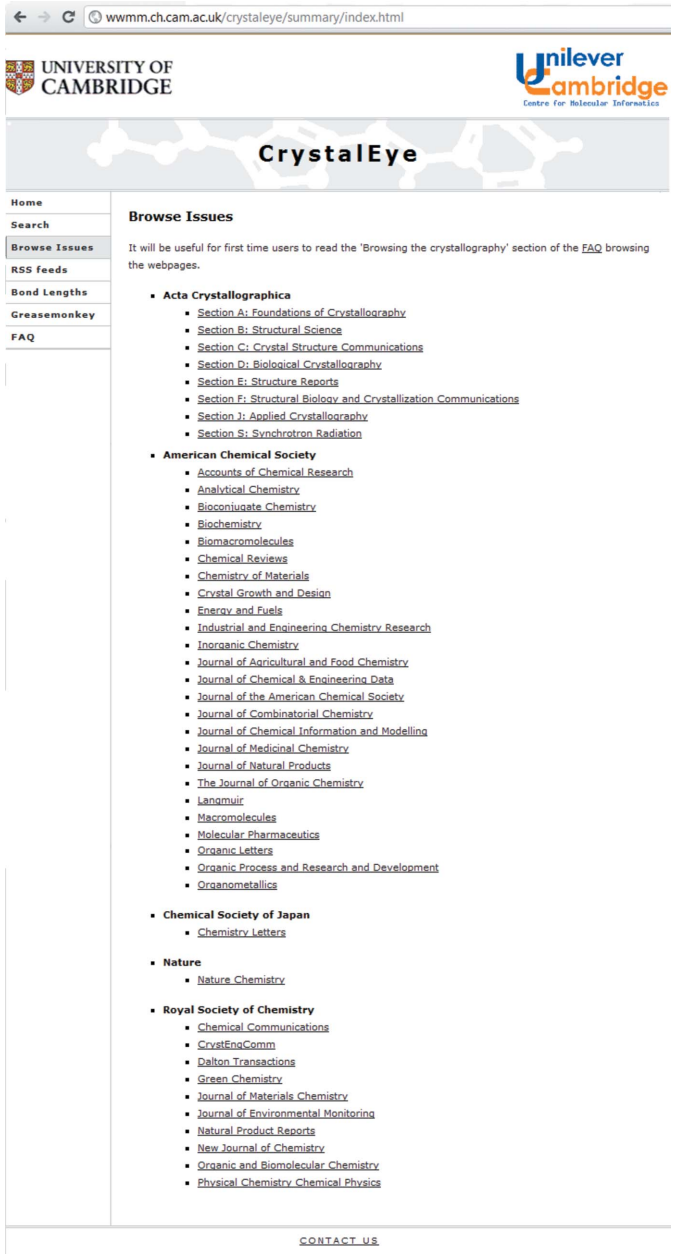

Figure 6
The complete *CrystalEye* list of publishers and journals that make CIFs freely available as supplemental data.

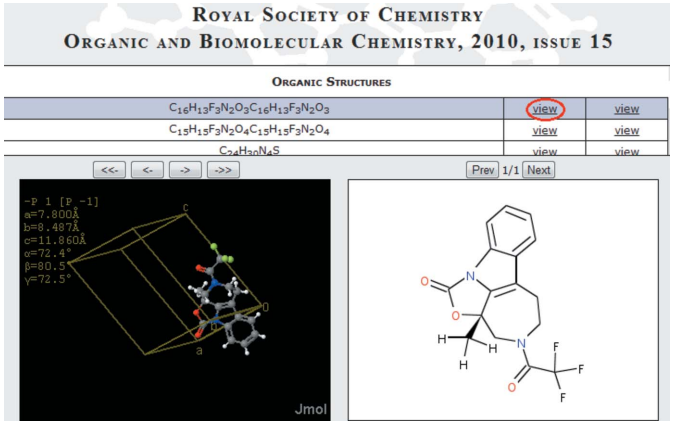

Figure 7
The first structure (highlighted) represents a derivative of the reactant (two-dimensional diagrams are auto-generated from connection tables using the CDK package and give a workable representation of the structure, including the absolute and relative stereochemistries). The article 'view' hyperlink (circled) routes us, *via* the DOI, to the primary publication.

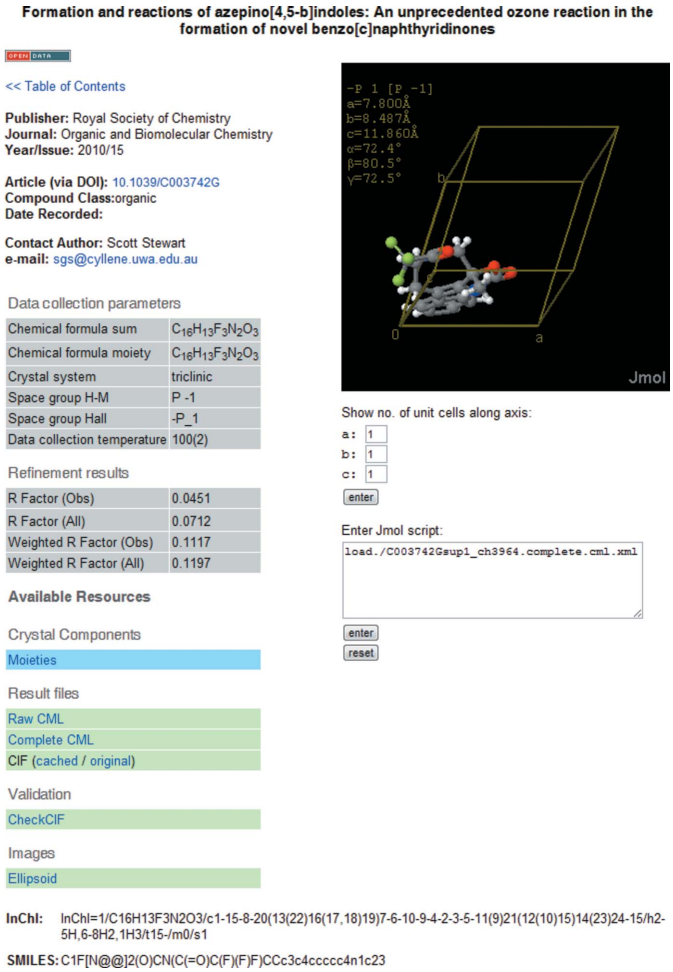

Figure 8
The summary page records the primary bibliographic data and a DOI linking back to the publication. The crystallographic data inform us that the structure is racemic (the space group $P\bar{1}$ has a centre of symmetry) and that the data collection temperature (100 K) means that the structure is likely to be of sufficient accuracy and precision to assert confidently that the molecule is represented. The crystallographic R factors are the primary reference for quality, and 0.04 would be regarded as indicating a structure of good quality.

3.4. The *CrystalEye* browser

A major part of *CrystalEye* is that it provides pre-computed, comprehensive, hierarchical browsing of key crystallographic and derived resources. At each stage of the hierarchy, the user knows the context of the molecule, data set, journal and publisher, and associated metadata such as authorship and date. We have found that, for many exploratory purposes, this type of browsing is superior to a complex set of web forms and queries where the new user has relatively little idea where to start. Every page is supported by the extremely powerful *Jmol* system, which allows users to re-orientate molecules, examine intermolecular geometry, explore surfaces and polyhedra, and generate crystallographic packing from the symmetry and unit-cell translations. In this way, we re-use the efforts of the Blue Obelisk community (http://blueobelisk.org/) and avoid having to duplicate a major crystallographic system. We illustrate the approach and architecture by four 'tours'. The first, illustrating browsing of journal articles, shows how we can start at a specific publisher and drill down to individual articles. Additional tours illustrating *CrystalEye*-derived data and indexing features (utilizing data from Galy *et al.*, 2002; Pascu *et al.*, 1998; Spinney *et al.*, 2007; Vickaryous *et al.*, 2005) are included in the supplementary material.

We start from the complete list of publications covered by *CrystalEye* (Fig. 6). As an example we take the RSC journal *Organic and Biomolecular Chemistry*, which deals primarily with organic compounds, and look at the first paper in issue 15 (2010) (Fig. 7).

This paper (Stewart *et al.*, 2010) reports a reaction and the authors have obtained crystal structures of both a derivative of the reactant and the product. These structures are reported in a single CIF with two separate `data_` sections (authors' numbering CH3964 and CH4651).

We can display the summary page for the reactant (Fig. 8) *via* the right-hand 'view' hyperlink.

Although the crystal structure (as opposed to the molecular structure) was not the primary goal of the authors, it can be generated automatically by the *Jmol* applet, and, for example, we can show the lattice repeated in two unit-cell translations along all three axes (Fig. 9).

We now turn our attention to the equivalent summary page for the product structure (Fig. 10).

A crystallographer or theoretical chemist might be interested in the variability of molecular geometry, which can be explored from

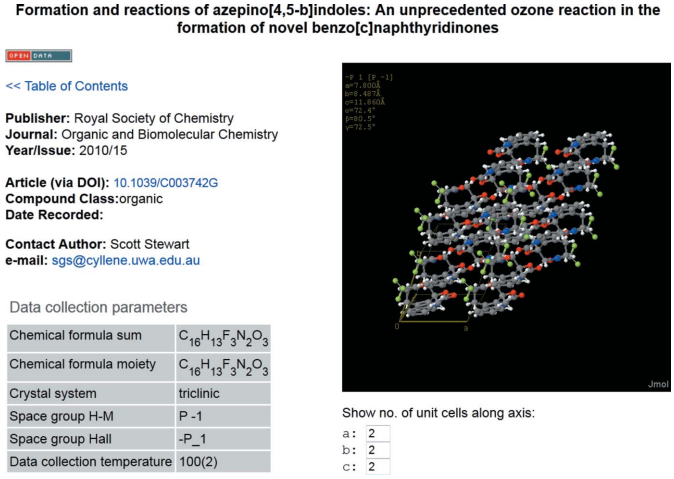

Figure 9
As Fig. 8, but displaying the lattice repetition functionality.

having three independent observations of the molecule in the same crystal. By following the 'moieties' hyperlink, the three independent molecules are presented to us, as shown in Fig. 11.

The chemical formulae of all three molecules are identical (see the structure diagrams in Fig. 11) but the geometries differ slightly. Each moiety has its own summary page (*via* the 'view' hyperlink), and by visual inspection we can see that the conformation of the hydroxymethyl group varies.

3.5. Enhanced journal web pages

Greasemonkey is an extension for the Mozilla *Firefox* web browser that allows users to install scripts into their browser that can make on-the-fly changes to HTML pages that the user visits, automatically adding or improving the information in the pages: so-called 'augmented browsing'. A *Greasemonkey* script (http://wwmm.ch.cam.ac.uk/crystaleye/gm/) based on Pedro Beltrao's script for highlighting articles (Beltrao, 2006), which automatically links articles to their *CrystalEye* records (Fig. 12), is available for download from the *CrystalEye* web site. Augmented browsing is an entirely client-side process, so completely under the control of the end user and independent of the content provider.

4. Comparison of *CrystalEye* with other systems

We are conscious that there are several other systems currently available for aggregating, analysing, preserving and disseminating crystal structures and derived data. These include the CSD, the Inorganic Crystal Structure Database (ICSD), the COD and other collections. While *CrystalEye* is not the largest or most comprehensive resource for crystallographic data, it has a number of unique points:

(*a*) It is completely open in that the data are taken from the public literature under sources where no intellectual copyright or other restrictions are claimed. A major concern with the CSD and the ICSD is that, although comprehensive, there are restrictions on their deployment and re-use, particularly regarding bulk processing and analysis. By contrast, the whole of *CrystalEye*, aggregated CIFs, bibliographic data and derived data (summary pages, fragments, bond lengths *etc.*), is completely open. *CrystalEye* has similarities to the COD and has ingested entries from this at various times.

(*b*) Immediacy. *CrystalEye* trawls the web every day and as soon as a journal issues new supplemental data it is immediately added to the *CrystalEye* system. We do not update every piece of derived data, and in particular the bond lengths and fragments have not been produced for recent structures. All new structures are transformed to semantic CML, and top-level summary pages and TOCs are generated. It is straightforward to add this information by running the software over recently aggregated CML files.

(*c*) We believe that the browsable approach to crystal structures and molecular components offers a complementary system to those developed with databases and search algorithms.

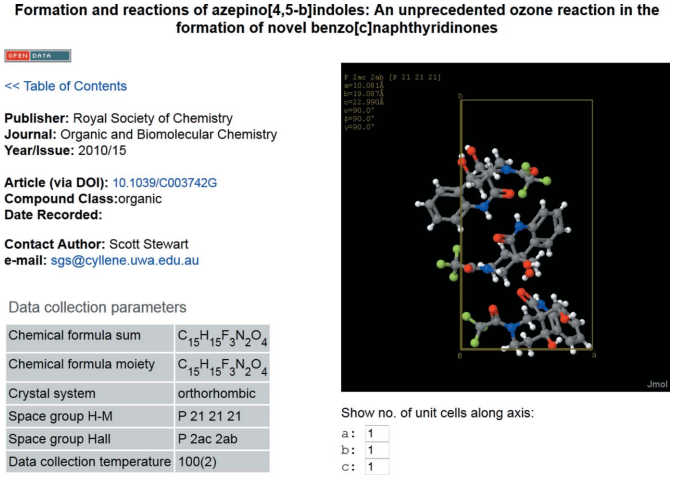

Figure 10
The product summary page. Note that there are three independent molecules in the crystallographic asymmetric unit.

5. Design and re-factoring

CrystalEye was not designed but evolved in response to a need. The current system has many excellent and useful features but has also suffered from 'first system complexity' (Brooks' law – 'plan to throw the first one away, you will anyway'; Brooks, 1975). In particular, the huge number of files generated and the unnecessary duplication of some components (*e.g.* ten million copies of the *Jmol* applet) require re-factoring.

The strong points of *CrystalEye* are the following:

(*a*) The automatic and comprehensive aggregation of public crystallographic data.

(*b*) The automatic transformation into semantic form and the addition of standard chemical structural information (connection tables, diagrams *etc.*).

(*c*) The ubiquity of links back to the primary literature and other data sources.

The derived data (moieties, fragments *etc.*) from *CrystalEye* will be preserved in static form but are not seen as an essential part of a *CrystalEye* system.

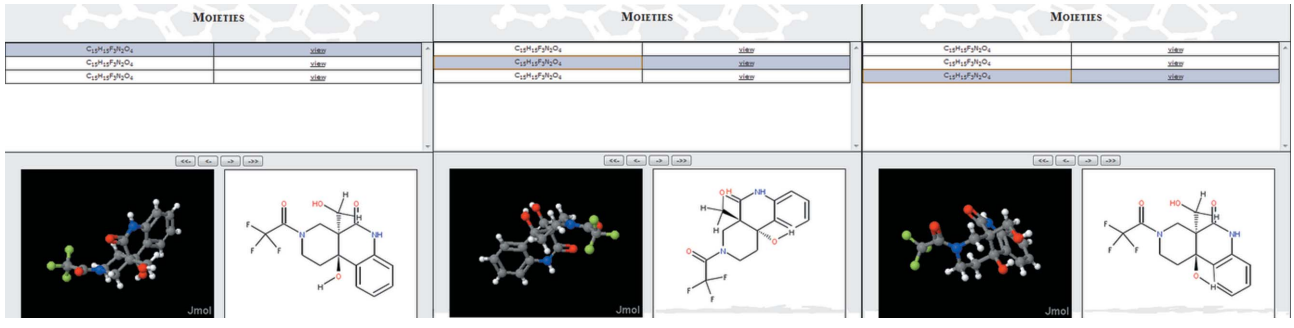

Figure 11
A composite diagram of the moiety overview page, showing the three instances of the product molecule.

Figure 12
The *CrystalEye Greasemonkey* script detects when the user is browsing a journal issue TOC page on a supported publisher's web site, and uses the *CrystalEye* DOI index to identify and highlight any articles in the issue that have *CrystalEye* reports, inserting links to *CrystalEye*'s structure summary pages directly into the journal TOC (*e.g.* [1] and [2] link to the summary pages on the left and the right of the figure, respectively). This makes it straightforward to identify those articles containing crystal structures in their freely accessible supplementary data, and provides the reader with quick and easy access to interactive three-dimensional models, saving them the effort of downloading the CIFs and loading them into a molecular visualization program, possibly needing to convert them into a suitable format along the way.

6. The future sustainability of *CrystalEye* and open crystallography

This article records the creation and implementation of the *CrystalEye* concept. *CrystalEye* is a project that has run at Cambridge for four years and has comprehensively crawled the open crystallographic literature for papers published in the period 1991–mid-2011. The accumulated material has been created as static HTML pages which can be mounted on any HTTPD-compliant web server. This resource consists of data for about 250 000 crystallographic data sets and derived information. The technology is openly available and stable and has been transferred to the IUCr offices in Chester. We believe that *CrystalEye* is a valuable static resource which, although inevitably not comprehensive, provides a rich crystallographic experience for browsing and searching.

At the IUCr 2011 Congress (Microsymposium 89), one of us (PM-R) announced that the COD and *CrystalEye* systems would be coordinated to provide interoperability. This could lead to a reduction of the work required to check CIFs and add other value. We would see the systems as continuing independently but exchanging and normalizing data, and re-using each other's code where appropriate.

Recently we have refactored and redesigned the *CrystalEye* system. The product of this – *CrystalEye2* – represents the technology that we believe is required for the crystallographic semantic web. The *Chempound–CrystalEye* repository system is capable of ingesting CIFs and carrying out all the described semantic conversions to complete CML. In addition, it produces Resource Description Framework representations of the data, which are exposed through a SPARQL endpoint, and provide much richer search functionality and enable integration with other sources of linked open data. *CrystalEye2* is not intended as a comprehensive index of the global open crystallographic content but as a technology that can be implemented by groups such as departments or publishers.

We thank the following for funding: EPSRC (NED studentship; Pathways to Impact Award), the IUCr and Unilever plc. for the Unilever Centre for Molecular Science Informatics. We acknowledge the invaluable assistance of Dr Charlotte Bolton in the production of this manuscript.

References

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *Acta Cryst.* B**35**, 2331–2339.
Beltrao, P. (2006). *Postgenomics Script for Firefox*, http://pbeltrao.blogspot.com/2006/05/postgenomics-script-for-firefox-i-am.html.
Bernal, J. D. (1965). *Science in History*. New York: Hawthorne Press.
Brooks, F. P. Jr (1975). *The Mythical Man-Month: Essays on Software Engineering*. Boston: Addison-Wesley Longman.
Day, N. E. (2009). *PhD thesis*, University of Cambridge, UK.
Day, N. E., Murray-Rust, P. & Tyrrell, S. M. (2011). *J. Appl. Cryst.* **44**, 628–634.
Galy, J., Enjalbert, R., Lecante, P. & Burian, A. (2002). *Inorg. Chem.* **41**, 693–698.
Gražulis, S., Chateigner, D., Downs, R. T., Yokochi, A. T., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P. & Le Bail, A. (2009). *J. Appl. Cryst.* **42**, 726–729.
Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.
Hey, T., Tansley, S. & Tolle, K. (2009). Editors. *The Fourth Paradigm: Data-intensive Scientific Discovery*. Redmond: Microsoft Research.
Pascu, S., Silaghi-Dumitrescu, L., Blake, A. J., Li, W.-S., Haiduc, I. & Sowerby, D. B. (1998). *Acta Cryst.* C**54**, 219–221.
Patterson, A. L. & Love, W. E. (1957). *Acta Cryst.* **10**, 111–116.
Spek, A. L. (1990). *Acta Cryst.* A**46**(Suppl.) c34.
Spinney, H. A., Korobkov, I. & Richeson, D. S. (2007). *Chem. Commun.* pp. 1647–1649.
Steinbeck, C., Han, Y., Kuhn, S., Horlacher, O., Luttmann, E. & Willighagen, E. L. (2003). *J. Chem. Inf. Comput. Sci.* **43**, 493–500.
Stewart, S. G., Ghisalberti, E. L., Skelton, B. W. & Heath, C. H. (2010). *Org. Biomol. Chem.* **8**, 3563–3570.
Townsend, J. A. (2007). *PhD thesis*, University of Cambridge, UK.
Vickaryous, W. J., Healey, E. R., Berryman, O. B. & Johnson, D. W. (2005). *Inorg. Chem.* **44**, 9247–9252.