DATA SCIENCE JOURNAL

Hall, S R and McMahon, B 2016 The Implementation and Evolution of STAR/CIF Ontologies: Interoperability and Preservation of Structured Data. *Data Science Journal,* 15: 3, pp.1-15, DOI: http://dx.doi.org/10.5334/dsj-2016-003

RESEARCH PAPER

The Implementation and Evolution of STAR/CIF Ontologies: Interoperability and Preservation of Structured Data

Sydney R. Hall[1] and Brian McMahon[2]

[1] School of Chemistry and Biochemistry, University of Western Australia, Crawley 6009, Australia
sydney.hall@uwa.edu.au

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

Corresponding author: Sydney R. Hall

The global application of the Crystallographic Information Framework (CIF) to the molecular structure domain has been successful over the past 20 years. It is used widely by molecular science journals and databases for submission, validation and deposition. This paper will give an overview of the CIF implementation, highlighting its particular successes and occasional failures. It will also recommend criteria for the application of an ontology-based data management system to any information domain. The paper will conclude with some details of the latest STAR data definition language and the importance of methods to the preservation of derivative data items.

Keywords: crystallography; crystallographic information file; data exchange standard; machine readable dictionaries; ontologies

1 Introduction

Humans have a poor record for the long-term retention of information and most prior knowledge is already 'extinct'. Moreover, surviving data records are often incomplete and unusable because of a lack of metadata. While data archiving should have improved in the electronic age, in some areas it has deteriorated because of the high turnover rate of storage media with rapidly changing technology. Hardcopy archives remain the most dependable long-term records for some information.

This paper will suggest ways that data and metadata can be better preserved. The crystallographic information framework (CIF) (Hall & McMahon 2005) protocol, which is a subset of the STAR File syntax (Hall 1991), will be used throughout as an exemplar, and details of the CIF data management approach, which has been over twenty years in the making, are given.

In particular the paper will describe a general approach to the preservation of metadata in any information domain. It will focus on ontologies in which metadata are recorded as attribute values for each unique data item. We use the term ontology in Gruber's (1993) sense of 'a formal specification of a conceptualization'. The approach we follow is particularly well suited to structured information based on a formal taxonomy or a well-defined hierarchical classification scheme. The paper will emphasize the importance of these ontologies to the overall retention of information and knowledge. Although CIF is widely cited in this paper to illustrate the ontology-based data management approach in a specific branch of molecular structural science, the same techniques may be applied to any information domain, but perhaps especially to those in taxonomic disciplines.

2 CIF: Information Interchange in Crystallography

The ability to access knowledge is an essential requirement of any field, and this makes the efficient exchange of information of fundamental importance. This is widely appreciated, but the definition and expression of electronic data remain poorly coordinated in science. This poses major obstacles to

efficient data interoperability. It was for this reason that the International Union of Crystallography (IUCr) in 1990 adopted a novel approach to the handling of data associated with the determination of molecular structures from X-ray diffraction experiments (**Figure 1**). This approach was based on a completely free-format file using the STAR File syntax (Hall 1991), which at that time was considered to be unusual. This became known as the *Crystallographic Information Framework* (CIF) (Brown & McMahon 2006).

While the CIF data exchange approach is now one among many *universal data language (UDL)* approaches for defining and handling electronic information, this was certainly not the case at the time of its adoption over 20 years ago. Even so, it remains the simplest and most efficient approach for expressing and exchanging the types of data commonly used in crystallography, and consequently has wide software support in the chemical and biological structure communities.

For a data exchange approach to be successful it has to be *efficient* and *flexible*. It must cope with the increasing volume and complexity of data generated by rapidly advancing technology. This makes data storage a minor consideration compared with extensibility and portability of data management processes. Most importantly, improvements in computing technology continue to generate new approaches to harnessing semantic information contained within data collections, and to promoting new strategies for knowledge management.

The basis for *any* information exchange process is agreed-to rules for making this feasible for the supplier and the receiver alike, *i.e.* the establishment of an exchange *protocol*. This protocol needs to be established at several levels. At the first level there must be predetermined ways that data (*i.e.* numbers, characters or text) are arranged in the storage medium. These are the organizational rules that define the *syntax* or the *format* of data. There must also be a clear understanding of the *meaning* of individual data items so that they can be correctly identified, accessed and reused by others. At an even higher level a protocol may also provide the

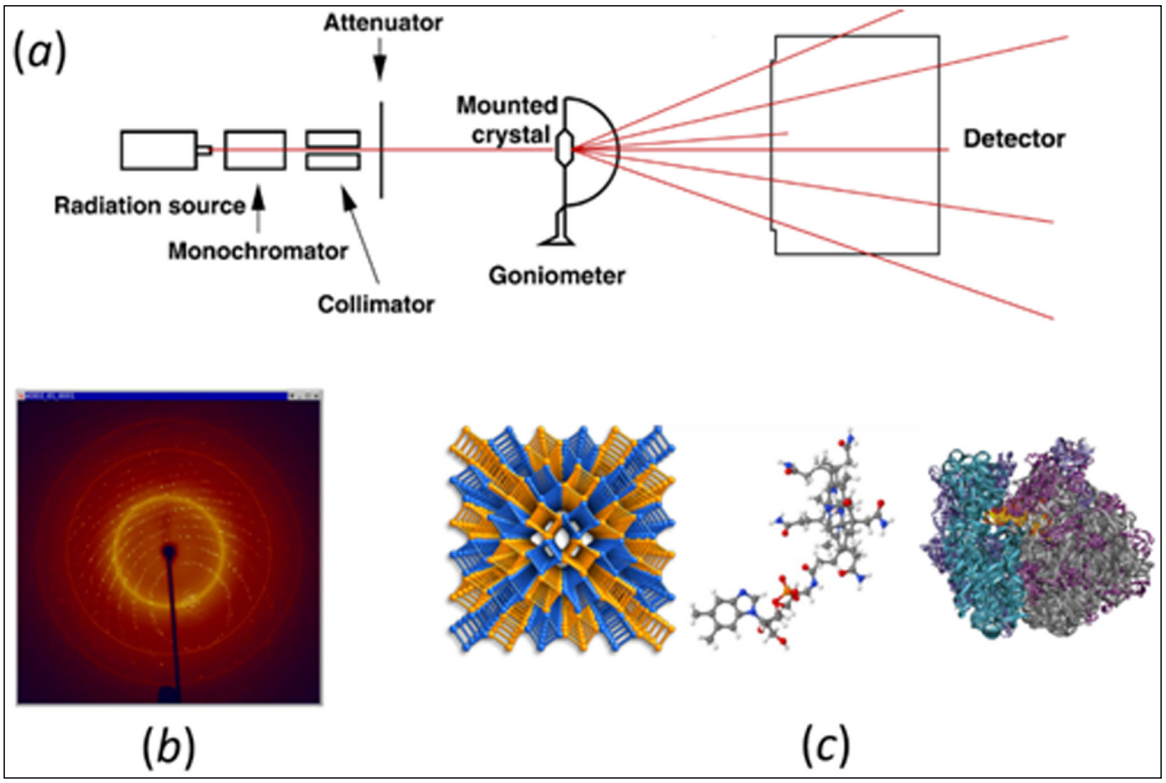

Figure 1: Determining molecular structure from crystallographic X-ray diffraction. (*a*) Schematic of a diffraction experiment. The apparatus may fit on a laboratory bench; or it may involve an intense X-ray beam from a synchrotron, a particle accelerator hundreds of metres in diameter. X-rays diffracted from regular planes of atoms in the crystal scatter into a number of beams, recorded as diffraction spots on an image detector (*b*). From the positions and intensities of the diffracted beams, atomic positions may be inferred and molecular structures deduced (*c*). The technique can be applied to inorganic materials, small organic molecules, large protein and nucleic acid molecules, or even viruses.

rules for expressing the relationships between the data, as this can lead to automatic processes for validating and applying the data values. These latter considerations provide the *semantic* knowledge needed for the rigorous identification and validation of transmitted and stored data.

For the CIF domain a comprehensive set of ontologies has been built for the disciplines of structural chemistry, molecular biology, powder diffraction, symmetry, charge density studies and several others (Hall & McMahon 2005). These ontologies define data items that are unique to their sub-domain. However, for reasons that will be discussed later in the paper, they also share metadata with a 'core' ontology containing the definitions of the most common items.

This paper has been divided into three major topics.

- A general description of an *ontology-based data management* approach, and how it provides data interoperability, particularly for taxonomic domains (§3).
- An *overview of how CIF was applied* in crystallography, and the data-handling benefits it has provided to the community (§4). This will take a critical look at the lessons learnt from over 20 years of CIF adoption and application.
- The latest extensions to the STAR language, and how these will benefit future ontology-based data management systems (§5).

3 Ontology-based Data Management

We structure the remainder of the paper as a list of responses to leading questions, in the style of a 'Frequently-Asked Questions' (FAQ) document. While the questions are designed to provide a logical development of our narrative, they do reflect the major concerns and queries that have been put to us over more than 20 years, either by individual software developers and data managers, or in invitations to contribute to a wide range of scientific, informatics, publishing and data management meetings.

The fundamental properties of an ontology-based data management approach, and its benefits to data interoperability, will be covered as answers to the following FAQs.

- What is the purpose of a data ontology?
- How do ontologies underpin interoperability?
- What are the critical issues when building an ontology--based management system?
- How is the decision to build an ontology initiated?
- What organizational support is needed to maintain such an ontology approach?

3.1 The purpose of a data ontology

Data ontologies record the metadata for data items used in a specific information domain. Ontologies are also referred to as *data dictionaries*, or, in XML, as *schema*. They are built using a particular adaptation of the data language, known as a *data definition language* (DDL). The metadata of each data item is specified in an ontology as a set of attribute statements. The attribute values describe particular properties of the item *e.g.* the *type* as a state value: *number*, *text*, *etc.* These values are the *metadata* of the defined item. The extent and precision of metadata has an enormous bearing on the understanding of the domain as a whole, and on how individual items can be applied and validated. Where possible every property of an item should be recorded in the ontology, including its relationships to other defined items. Indeed, these relationships, when expressed as mathematical script values for the *method* attribute, become an important resource for many ontological applications (*e.g.* automatic data validation) and provide the underpinnings for artificial intelligence and automata processes. Note that the DDL attributes are themselves defined in a domain-*independent* ontology: for STAR this is known simply as the DDL ontology; for XML, the *schema schema* (Thompson *et al.*, 2004).

3.2 Ontologies underpin interoperability

As discussed earlier, interoperability is the primary goal for a universal data language, of which STAR, CIF and XML are members. Interoperability allows data instances to be readily understood, exchanged and manipulated in a domain, without translation. It should be noted, however, that the data language used to express *data instances* (*i.e.* sets of data values identified by unique tags) conveys no metadata. These instances contain no usable information on the data values – the tags are purely for identification and have no semantic content. The *understanding* of values in a data instance depends solely on

the ontology that defines these items. Ontologies are therefore an essential adjunct to the processing of data if interoperability is to be assured.

In past handling practices, the knowledge of data items was embedded either in the minds of users or as code in computer programs. Such transient practices inevitably led to misunderstandings, mistakes and loss of information.

3.3 Critical issues when building an ontology-based management system

It is a commonly held belief that the critical first step, when planning an ontology-based data management system, is the choice of a file structure for expressing instance data. The CIF experience suggests that this is not the case. In fact, provided that instance data is expressed in one of the compliant UDLs, this choice becomes a relatively minor one compared with the level of commitment and cohesion needed by a discipline to build a data ontology. Such a commitment is a major one and must be preceded by a careful assessment of the benefits that will accrue to a discipline from the build. Notwithstanding the enormous definition advantages of an ontology for any taxonomic discipline, the need to develop a fully *interactive* ontology, which includes methods, will depend strongly on the nature of the archived data, and in particular, the extent of retained *derivative* information. The critical issues leading up to this commitment are essentially 'what are the needs of the data community', and 'who should administer the ontology development'.

The Need to Build

A dominant factor in the success of an ontology implementation is the extent to which large-scale data *exchange* is actually required within a community. In some disciplines data may be expressed using relatively simple metadata, and a data archive may be considered to be essentially an inventory. For example, in botanical specimen archives where the majority of data items are primitive (*i.e.* observed or measured), and the data relationships to derivative information are relatively few and simple, an interactive ontology is less critical. Provided that a user community can readily comprehend the meaning of the archived data (in the case of botany, the characteristics of entries in a specimen collection), then a rather simple spreadsheet or similar *ad hoc* format may suffice for many search and recording purposes. Where a full ontology becomes important is when users need to critically validate or re-evaluate individual data points, or where it is necessary to test what data are invalid, inaccurate, incomplete or perhaps generated by subtle effects not apparent in the raw (*i.e.* primitive) data values. In such instances there is likely to be a need to understand complex dependencies between data items, and then richer metadata are required. This is where interactive ontologies can really extend and strengthen the meaning of the data, and of the scientific knowledge that they represent.

The Focus of a Build

When a decision has been made to proceed with an ontology-based management system, we recommend that the main focus be on the ontology construction, not on the choice of UDL for representing instance data. If discipline practitioners have experience with a particular UDL (*e.g.* XML), then their familiarity with tools that support this UDL might be beneficial. On the other hand, where they do not already have such expertise, they might find it easier to prototype their ontologies using the comparatively lightweight STAR File syntax. Early decisions on which DDL attributes are appropriate for the domain, and their application to the definition of needed data items are where the most effort is needed, and where most of the gains are to be made.

The Builders

Of equal, or perhaps of more, importance to the commitment to build a data ontology is the need to involve and empower authoritative institution(s) in the domain who represent the majority of practitioners. These organizations need to give strong support to the ontology-based approach if it is to succeed in the long term. Building domain ontologies is a big undertaking and can involve the efforts of the best data experts in the field for several years. The magnitude of the commitment needs to be understood in advance and the choice of an organization with both the authority within the discipline, the administrative capability and the long-term incentive to coordinate this may be quite limited in some disciplines. For the CIF implementation, the crystallographic community was fortunate in having a strong, forward-thinking international body in the IUCr, who recognized a global ontology as benefiting in particular its journal operations, and those of the major crystallographic databases. Indeed the major beneficiaries of a global ontology are *journals* and *databases*, and that begs the question of whether these organizations

might for some disciplines be the most appropriate for the building and maintenance of a data ontology. This will be discussed further in Sections 3.4 and 3.5.

The Benefits

Notwithstanding the considerable work needed to build an ontology-based management system, the effort is relatively small compared with the enormous benefits that accrue to a discipline when it is complete. However, the data handling gains only come when all of the basic ontology machinery is in place – in the interim, the developmental effort will certainly test the commitment of the discipline. Indeed, during this gestation period the task can seem almost counter-productive. Having participated in an ontology design and implementation, we can say that the final benefits to archival and publication processes are really worth the large effort!

3.4 How is the decision to build an ontology initiated?

Typically a decision to proceed down the ontological route is initiated by the people or organizations in charge of data management within the discipline. In the crystallographic domain, these are mainly represented within the IUCr Commissions on Data, and on Computing. Somewhat unexpectedly, however, the CIF initiative actually came from a resolution at the 1987 IUCr Congress that IUCr *journals* should be made electronic as soon as possible. This led to an interim Working Party being formed, and their recommendations on a new crystallographic information file approach was proposed and adopted at the 1990 IUCr Congress.

The implementation and uptake of CIF was rapid, being driven primarily by a desire to make the IUCr journal submission and production processes electronic as quickly as possible. Inevitably the short deadlines set for this implementation led to inadequate community involvement and consultation in some aspects, and this definitely required some remediation later when launching some of the CIF journal applications. Nevertheless, the IUCr, as the authoritative international body representing the domain, was fully committed to this development, and as alluded to above, this was absolutely essential for such a major undertaking.

We consider it unlikely that the commitment and authority needed to build an ontology which is setting *global* data standards can be provided by a commercial organization. Moreover, the magnitude of the initial data conversion tasks for new data standards, as well as the long-term effort required to maintain these standards, strongly militates against such a development being initiated by a non-representative body. The success of any ontology implementation depends critically on the involvement of *all* stakeholders (scientists, journals, databases, manufacturers and software developers) especially during the application phases, and the CIF ontology development received, with the backing of the IUCr, that support. And, because the primary motivation for the CIF development was the enhancement of journal and database activities, the major stakeholders in these activities for the discipline, the IUCr, the Cambridge Crystallographic Data Centre (CCDC) and the Protein Data Bank (PDB), saw universal standards for data exchange as an unquestioned common goal.

The CCDC and PDB structural databases are organizations that started within academic research groups and grew to serve the community as a whole. The CCDC was an early and enthusiastic contributor to the CIF ontology because it provided an efficient means of ingesting structural data from both journals and individual depositors. At the inception of CIF in the early 1990s, the CCDC contained nearly 100,000 entries (it now holds over 800,000), and efficiency and accuracy in adding new structural data as CIFs has contributed to its growth and success. Prior to CIF, the database exported structural data in its own standard format but direct re-use was relatively uncommon. The ability of CIF to describe measured experimental data in addition to derived structural coordinates may go some way to explaining its greater success as an exchange format.

Somewhat in contrast, the macromolecular structure community, which deposited coordinates with the PDB, found it very useful to exchange and re-analyse protein structures using the in-house PDB file format. In the early 1990s the PDB held a few hundred structures (nowadays it has over 110,000), so volume issues were less acute than for the CCDC. However, there were already limitations with the PDB format, and the growing database was re-engineered with a relational schema that mapped one-to-one with the CIF ontology developed specifically for macromolecular structures (mmCIF). In spite of this, the community continued for many years to use the old PDB format for exchange of coordinates, and mmCIF is only gradually becoming the working format for the latest generation of software. In this case, the existence of a functional, if restrictive, exchange mechanism (and software that supported it) acted as something of a brake on the uptake of mmCIF.

It is worth noting that within the crystallographic domain it *might* have been possible for either the CCDC or the PDB to become the primary ontology builder and maintainer if the IUCr had not taken on that role, but in our view it is unlikely. These large databases remain active in, and supportive of, the CIF development (for example, the PDB acts as the agency by which the structural biology community extends the mmCIF ontology) but they each have a specialized focus that would most likely have resulted in different data standards within different subsets of the community. In non-crystallographic disciplines, it may be entirely possible that a large database or journal group has the capability to be the primary developer of ontology-based data standards. But a successful long-term implementation will hinge critically on wide support within the user community, and on the ability of the organization to commit to the financial and administrative effort that is needed for any global ontology development.

3.5 What organizational support is needed to maintain such an ontology approach?

Because ontologies are time-dependent and extensible in nature they require significant ongoing maintenance, and this needs to be coordinated by a key organization within the domain. In the case of CIF, an appointed Committee for the Maintenance of the CIF Standard (COMCIFS) controls all additions and modifications of crystallographic ontologies on behalf of the IUCr. CIF material and other related material is available on the IUCr website. Financial and perhaps personnel support is needed to maintain ontologies and ensure their wide promotion. This will probably involve: staff within the organization to assist the activities of the maintenance group and to provide web updates; the funding of workshops for authors/users/programmers; as well as possible seed funding for software developments.

There is also a question of how much in the way of resources an organization is prepared to commit when there are competing demands for support. In the chemistry world during the early years of the Millennium, the International Union of Pure and Applied Chemistry (IUPAC) was actively considering support for Chemical Markup Language (CML), an XML-based ontology of chemical structure with projected extensions to spectra and reactions (Murray-Rust & Rzepa 1999), and for InChI, a deterministic chemical structural identifier scheme (Davies 2002). In the event, InChI was supported and has achieved widespread use and application, while CML remains a dormant standard. We are not sufficiently close to IUPAC to make any value judgement on this, but we suspect that the relative success of InChI arose from a more immediate need for sharing a specific type of information, in line with our remarks above that the greater the need for data exchange within the community, the more impetus there will be for standards to be adopted. If this observation is true, the current growth in funding bodies' requirements for data sharing may well provide an incentive for new ontologies to be developed or for existing ones to be reinvigorated.

It also suggests that, whatever level of organizational support might be mustered, the rate of uptake of an ontology data management system will be driven more by bottom-up community requirements than by top-down design ideals.

4 Overview of CIF Implementation

The successful implementation of the CIF data management system over 20 years provides a useful exemplar for similar approaches in other domains. Some aspects of the implementation could be improved on and these will be discussed. The topic will be given as answers to the following FAQs.

- How was *data* expressed before CIF?
- What was the implementation sequence?
- What were the hurdles for CIF adoption?
- What are the main benefits of CIF?
- What is the impact of CIF: good and bad?

4.1 Data expression approach before CIF

The field of crystallography has many sub-disciplines; some can trace their origins back to prehistory, if one includes perhaps the practicalities of Bronze-Age metal working or the cleavage and mounting of ancient Egyptian jewellery! However, most recent activity in this field revolves around the imaging of molecules using the X-ray diffraction data of crystals (see **Figure 1**). When these experiments started early in the 20th century diffraction data was measured visually and the results were recorded on paper. Such practices continued unchanged up until the advent of computers in the mid-1950s. All this time the experimental results were recorded and sent for publication as handwritten or typed documents. With electronic computers came the use of paper tape or Hollerith cards for data storage, but publication approaches remained largely

unchanged. Also, throughout these times, raw diffraction data were rarely archived or published. From 1965 onwards magnetic tapes were preferred for long-term data storage but until 1985 publication approaches changed very little. Regrettably many journals published only the coordinates of the atom sites and/or a graphical representation of the molecule, and some were reluctant to publish even atomic coordinates. Most did not ask for, or save, diffraction data!

The rapid growth of the internet, and its ability to globally transmit large amounts of information, completely revolutionized data exchange. For journals and databases the main data hurdle was now the large variety of formats in active use. In the early 1980s the IUCr promoted, through its Commissions on Data and Computing, the development of a standard fixed format for chemical molecular structure data. This resulted in the Standard Crystallographic File Structure (SCFS), a fixed-format approach, which was released in several versions, the last in 1987 (Brown 1988). The SCFS approach was used and promoted by the IUCr journals and was implemented by several software packages. However, it was not widely accepted for one important reason – its fixed format was quite inflexible to the addition of new data items. All sciences, and for that matter all human endeavours, are continually changing, and the need to include new information is an intrinsic requirement for any data recording process. The SCFS approach was incapable of this without serious modification and a subsequent incompatibility with prior data. SCFS therefore had to be replaced by a storage approach that was extensible.

4.2 What was the CIF implementation sequence?

The main factors leading to the adoption of CIF in 1990 have been alluded to above. The proliferation of formats made the transmittal of data to journals and databases unduly complicated. This was relieved briefly by the SCFS format, but this approach was non-extensible and this was an unacceptable constraint. A motion was passed at the 1987 IUCr Congress that IUCr journal processes (submission and publication) should be made electronic as soon as possible. This led to the formation of a Working Party on Crystallographic Information (WPCI) headed by Ted Maslen (University of Western Australia). In the following year, at the European Crystallographic Meeting in Vienna, the WPCI proposed that the IUCr adopt a free-format file approach based on the STAR File syntax (Hall 1991). As a consequence, a working group was formed to adapt the STAR File syntax to crystallographic data and applications. Discussions were held with representatives of major software packages, equipment manufacturers, journals and databases, and a syntax based on STAR was proposed. This syntax restricts data lists to a *single nesting* (STAR has unlimited nesting) and excludes *save frames* (STAR addressable data cells) from data instance expression. The approach was named the *Crystallographic Information File* (later changed to the Crystallographic Information *Framework*). The CIF proposal was put to the IUCr Executive and General Assembly at the 1990 Congress, and accepted as a data standard. The CIF specifications, and the initial defined data items, were published in 1991 (Hall, Allen & Brown 1991). The timetable for this implementation is given in **Table 1**.

Even prior to the publication of the CIF specifications and the data items to be used initially with CIF, development commenced on application software and a promotion of this approach. This was necessary for the IUCr journals office, for which CIF was principally developed. Most of the initial data items were those intended principally for IUCr journals (including text items for manuscripts); however, they were also suitable for the various curated databases of crystal structures already established. Programmers and equipment manufacturers also were encouraged to adapt their data generation software to these standards so that users could submit results to journals without major modification. Often the very existence of particular CIF items acted as an incentive for stakeholders in the analysis chain to satisfy the journal submission requirements. Prior to CIF the data output by packages was not always aligned to these requirements.

4.3 What were the hurdles for CIF adoption?

Whereas the uptake of the CIF approach by stakeholders and the community was on the whole relatively smooth, some aspects of the adoption were decidedly more difficult than others. It will be useful to those embarking on this journey to know what these were.

The first hurdle was somewhat predictable. In 1991 the most common computing language in crystallography was (and perhaps still is!) Fortran. Most programmers had little or no experience in parsing free-format data, all data then being in fixed format files. This gap was closed with the circulation of several STAR-compliant data checking routines (Hall & Sievers 1993; Hall 1993a; Hall 1993b; Hall 1993c) coded in Fortran 77. This code provided tools and examples for programmers who needed them. Many did not, however, because their software only *generated* CIF-compliant data. It is important to be aware that software

1987	IUCr Working Party on Crystallographic Information (WPCI)
1988	WPCI recommends STAR File as data format
1989	CIF syntax derived from STAR File approach
1990	*Acta Crystallographica Section C* recommends CIF use
1991	First CIF data and text submission
1992	*checkCIF* validation facility introduced
1994	CIF web upload and auto-validation
1996	*Acta C* also publishes papers authored solely in CIF format
1996	*Acta C* mandates CIF-format submissions only
2001	*Acta Crystallographica Section E* online: short structure reports in CIF format
2005	*Acta Crystallographica Section F* online (short reports of macromolecular structures); *International Tables for Crystallography* Vol. G published

Table 1: Timeline for CIF implementation by IUCr journals.

vendors and developers are usually not enthusiastic about global data formats. These can cause additional work and disruption. For this reason programmers usually prefer users to work with *their* formats. Global standards enable users to migrate easily to other software packages and that is definitely not the preferred business model! It follows that with the introduction of a new data standard there needs to be real incentives to get programmers quickly involved. In the case of CIF the 'carrot' was the new *Acta C* data submission requirements issued in 1992.

Another potential hurdle for any new data model is the need to educate the community. A training programme promulgating the new syntax must be part of the adoption process. Early in this process forums were organized to give instruction on how to correctly compose and modify CIF data. Because the tag–value syntax of CIF was considered straightforward, indeed, simpler than most formats then in use, this step was not a priority during the earliest CIF implementation. However, the new format did prove difficult for some; they failed to appreciate that every value needed an identifying tag, and that the correct spelling of a tag is essential to its recognition. The education gap was closed over time through a combination of journal workshops, pamphlets, and direct assistance by journal staff. Eventually the problem was almost completely alleviated with the availability of CIF editors that detected syntax violations (*e.g.* Allen *et al.*, 2004; Westrip, 2010). The take-home lesson here is that the time it takes for a user community to become familiar with a new standard is not short, and that significant education is needed from the outset to bring all users onboard.

The need for compliance training applies to more than just generating instance data files for submission or deposition processes. A UDL supported by an active ontology is a coupling that provides important new avenues to the intelligent and automatic processing of instance data. The validation of data values in an instance data file against those expected from relationships with other data, is a case in point. Many journals now utilize the IUCr *checkCIF* facility to validate publication submissions. These checks include syntax compliance; inadequate precision of items; as well as the inter-item consistency of values. The results of these checks have become an essential part of the editorial and acceptance stream. In many respects, *checkCIF* has revolutionized the refereeing process and reduced overall publication times. However, some authors found this level of automation difficult, and there was some resistance to the reduction of author–coeditor exchanges. Of course, there is likely to be some opposition to major changes in any endeavour – anticipate this reaction and be prepared to promulgate the fact that the quality of processes is being enhanced, being made more consistent, and much faster.

These benefits will be obvious to most working in the data field but some users will take more convincing, and others are prepared to forgo these advantages so they can keep doing what they have done in the past! This means that prior processes may have to be kept operational longer than first thought. The well-understood maxim here is that achieving any data paradigm shift depends more on the rate of community absorption than the time needed to change the supporting technology.

Another matter requiring early care in the implementation process is the need for an agreed-to process for adding new data items. New items should conform to the hierarchy and organization of the ontology and must be carefully moderated if they are introduced to a data management system. Anyone can place their own tag–value items into a CIF, but until their definitions are accepted into the global ontology they can only be for local use. For CIF items, acceptance into the official global ontology is the responsibility of the

IUCr COMCIFS group, which is composed of data experts from various sub-fields. COMCIFS ensures there is no duplication of the tags and that they fit into the IUCr tag-naming scheme. This committee is also responsible for ensuring that the ontology definitions are correct and up to date.

4.4 What are the benefits of adopting CIF?

Most of the advantages accruing from the CIF approach have been discussed in the earlier sections. Its extensible syntax enables new data items to be added without corrupting existing data archives. The most important benefit of CIF is of course its interoperability. All journals and databases in crystallography read CIF data, as do many software packages and graphics tools.

The success of CIF within its domain illustrates the accumulation of benefits as a standard comes to be accepted within a community, though we must emphasise that community-wide buy-in does not necessarily happen overnight. As mentioned previously, the original target of CIF was the small-molecule structure reports arising from single-crystal diffraction experiments, published in a number of crystallographic journals, and ingested by curated databases of organic, inorganic and metal structures. Subsequent granular ontologies (known as CIF dictionaries) were commissioned for use by databases recording biological macromolecular structures (Protein Data Bank and Nucleic Acids Database) and structures characterized by powder diffraction (International Centre for Diffraction Data). Other, smaller dictionaries followed to describe specialized areas of study (*e.g.* multipole electron density, or restraints applied during least-squares structure refinement calculations); a full list is given below in **Table 2**. More recently, a detailed dictionary has been produced that describes the raw image data collected by the experimental instruments, and allows the user to couple the image data *in the same format* to the operating conditions of the instrument and sample – information that is usually considered in other approaches as 'metadata'.

In principle, therefore, a complete pipeline exists (**Figure 2**) from experimental data collection, through data processing and reduction, structure solution and refinement, annotation, validation, publication and archiving, in which the same file could be carried along, progressively growing as further information and content is added to it. In practice, of course, this is not actually done. Raw images are not required by journals, partly on account of their size (a collected data set for solving a crystal structure can be several gigabytes), and so are stored separately at synchrotrons or by individual laboratories (or discarded). The reduced data sets actually used in the solution of the crystal structures, which are required to be deposited with IUCr journals, may be uploaded as distinct files from the final atomic positional coordinate data (even though they are in the same format).

As mentioned previously, the file format or, more generally, choice of universal data language (UDL), is not critical for the success of the ontological approach to data management. It is, of course, helpful

Crystallographic Core	*Shared with all crystallographic ontologies*
Crystallographic Diffraction Image	
Crystallographic Macromolecular Structure	
Crystallographic Modulated Structure	
Crystallographic Powder	
Crystallographic Multipole Density	
Crystallographic Restraints	
Crystallographic Symmetry	
Crystallographic Twinning	
Data Definition Language 1	*Used in crystallographic core, modulated structure, powder, multipole density, restraints, twinning*
Data Definition Language 2	*Used in crystallographic macromolecular structure, diffraction image, symmetry*
Data Definition Language 3	*CIF version of DDLm being applied to crystallographic dictionaries*
Data Definition Language m	*'Methods' language with STAR2 specifications published in J. Chem. Inf. Model. (2012) [19]*
Nuclear Magnetic Resonance	
Molecular Information File	
QCHEM: Quantum Chemistry	

Table 2: The ontologies currently constructed with STAR DDLs.

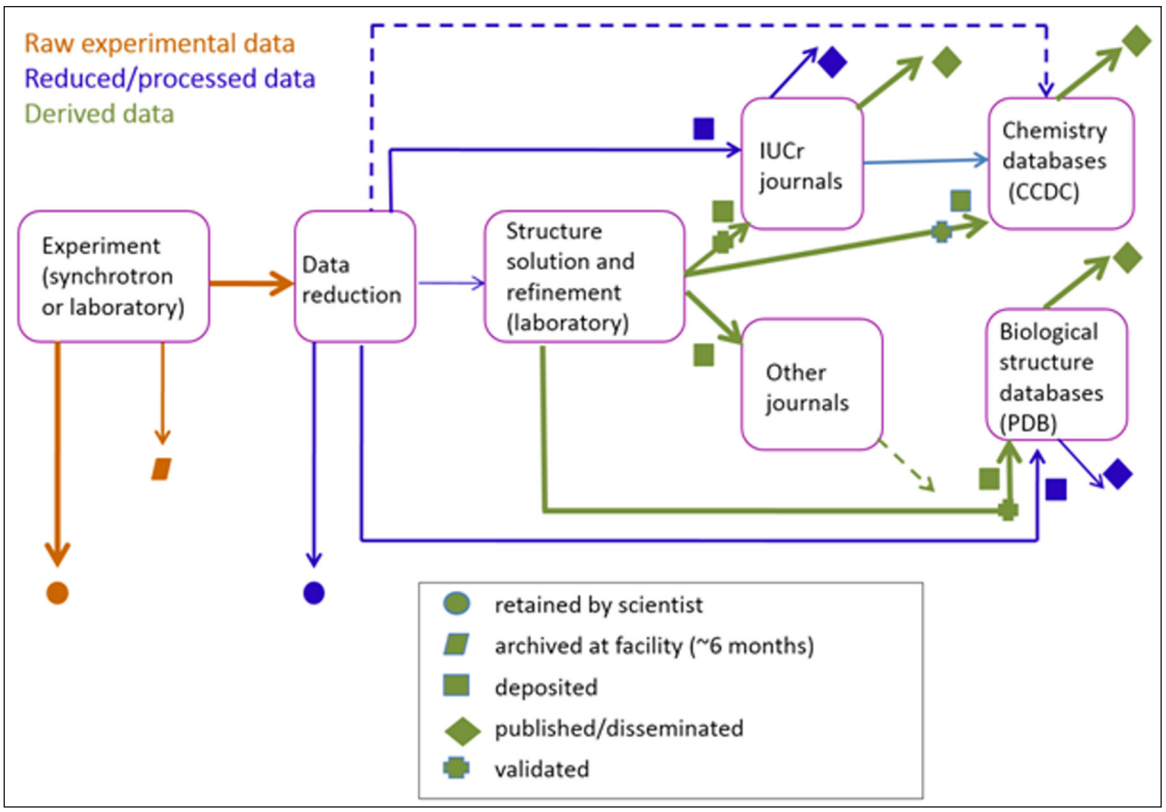

Figure 2: A coherent information flow in crystallography. CIF ontologies characterize data at every stage of the information processing life cycle, from experimental apparatus to published paper and curated database deposit.

within a single domain if a single UDL choice is widespread, and CIF dictionaries do ensure that crystallographic information across a very wide range of subdisciplines can be handled with the same software tools. However, even within crystallography, different UDLs are sometimes used. The Worldwide Protein Data Bank, which is a worldwide consortium of participating groups across three continents, has developed an extensive software system to handle macromolecular data using the mmCIF ontology, but instantiated in a number of different formats; and it uses XML to transfer crystallographic data internally, in order to facilitate bulk digital transmission and validation (Westbrook *et al.* 2005). The use of 'imgCIF' (the particular format describing diffraction images; Bernstein & Hammersley 2005) is declining as detectors become more powerful and have greater data collection rates than can easily be accommodated within the CIF format. Nevertheless, the current working file format (NeXus) has tokens that are derived from the imgCIF ontology, so that it remains possible to translate losslessly between NeXus and imgCIF formats.

In all these practical applications, the characterization of data on the basis of CIF ontologies allows any specific data item to be located at the appropriate point, or points, along the information transfer pipeline illustrated in **Figure 2**. So far as we are aware, no other experimental science has a comparably complete ontological framework. We like to characterize this pipeline as providing a *coherent information flow* from instrument to publication and beyond (another application of that most useful acronym, CIF!).

Perhaps the greatest benefits of CIF are yet to be realized; its latest ontologies contain a much richer level of metadata, including detailed method scripts. Most of the existing validation applications, such as *checkCIF*, do not yet use these ontologies but rely more on internally coded knowledge. The next generation of CIF software, of which *JsCif* (du Boulay 2013) is an exemplar, will use metadata in ontologies as the knowledge base for advanced evaluation, validation and definition techniques. The experience gained in one field can inform definition and manipulation of metadata by practitioners in any field (Hester 2015).

4.5 What is the impact of CIF: good and bad?

The single most important impact of CIF has been the ability to easily access data globally across journals, databases and laboratories. Portability provides significant efficiencies to submission–deposition processes which are then simpler and faster, and to the clarity of data exchanges that are enhanced by using a common

data language. Another important impact is that the ontology definitions have led to a much better understanding of particular data items across laboratories and databases. For some items the process of reaching agreement on their ontology definition has led to a new way to express relationships between items. It was also discovered that some software had for a long time been using different algorithms to calculate commonly used derivative data items; these differences were detected and resolved through agreement on the ontology definitions.

Although relatively few in number, there are also some less positive aspects of CIF. The fact that CIF and its ontologies need to be mediated is certainly seen by some to inhibit rapid growth of global data portability. These critics would advocate a less constrained data regime, so that advances in technology can be more quickly reflected in new data items and application software. It is an approach often associated with the 'semantic web' and 'open source' developments. We think it is correct to say, however, that in the area of knowledge interoperability none of these efforts have achieved what the CIF development has. In fact, a recognized guru in the semantic world has suggested (Murray-Rust 2014) that CIF is '*the best working data system in the scientific world*'. It is clear that a reasonable level of mediation is needed for any data management system if it is to be a dependable global standard.

5 Evolution of STAR Ontologies

The STAR File syntax underpins the CIF application in the crystallographic domain. CIF has been tailored to the nature of crystal and molecular data, whereas the STAR language has the capacity to handle the more complex data representations found in disciplines such as chemistry, quantum chemistry, botany, *etc*. Moreover, the syntax of STAR can expand to meet the changing data needs in fast evolving disciplines. Needless to say, extensions to the STAR syntax maintain strict compatibility to all existing STAR data instances, including those in CIF. This topic will be covered as answers to typical FAQs posed by data managers interested in using STAR.

- What are the current applications of STAR?
- Why does STAR syntax need to evolve?
- Why is a record of data origin important?
- Why should domain ontologies be shared?
- Why is exact data relationship information important to ontologies?

5.1 The current applications of STAR

By far the most important application of STAR to date has been the CIF management approach within the crystallographic domain. Details of this implementation have already been given. The CIF and other ontologies built from STAR are shown in **Table 2**. It is worth pointing out here that, in addition to the expression of instance data, discipline-specific ontologies are expressed in DDL protocols and attributes that also fully conform to the STAR syntax. This means that the algorithms used to parse instance data are identical to those for ontologies.

STAR has already been applied to some non-crystallographic domains, and others are being considered. Important to all these applications is that the information contained in a STAR File will be identical when using another widely-used UDL, the XML file. Software (Spadaccini, Mildenhall & Hall 2008) already exists for moving data between these two UDLs but it does depend on the existence of conformant ontologies.

Of the non-CIF STAR applications, that for the NMR imaging field (*Biological Magnetic Resonance Data Bank*) is the most supported. This is a field closely allied to diffraction imaging disciplines using CIF, but requires nested data lists which meant the CIF approach was inappropriate. Another STAR application, the Molecular Information File (MIF) has been published (Allen, Barnard, Cook & Hall 1995) but has not found significant support in the chemistry community. As already discussed, data standards must have wide support to be an effective data management approach. The *QCHEM* ontology for quantum chemistry (illustrated in Spadaccini *et al.*, 2005) is another STAR creation that has been in need of discipline support, and recently appears to have found a new champion. We have also carried out some work on botanical ontologies (*Florabase*), which has provided useful insights into how they differ in small but significant ways from those in crystallography. For example there is a need to use Boolean relationships between the enumeration states of particular items. In addition, plant naming hierarchies lend themselves to the use of 'definition' methods to connect these hierarchies.

It cannot be repeated too often that the success of ontology applications depends more on the involvement of experts in the domain than on the nature or sophistication of the data language!

5.2 Why does the STAR syntax need to evolve?

This need has already been alluded to. Every information domain is subject to continual change. These not only require the addition of new data items, but need new ways to express and organise this data. A UDL must be able to respond to change in a way that enables new data to be conveyed in a coherent and intuitive way. Syntax changes must be backwards compatible. Recently a number of extensions to the STAR syntax were published (Spadaccini & Hall 2012a). These extensions increase the allowed characters to embrace the complete UNICODE set. This permits multi-language characters to be used in tags. Other extensions introduce a wider range of multi-line token delimiters, including the [ ] and { } braces for lists, arrays and tables. The latest STAR syntax has a more explicit protocol for addressing *save frames* using a reference-value encoding table (*Ref-table).* All of these extensions are needed to satisfy data uses that did not exist in 1991. They have already been applied in the formulation of a new data definition language DDLm (Spadaccini & Hall 2012b) as well as in the script language *dREL* (Spadaccini, Castleden, du Boulay & Hall 2012) for the DDLm method attribute.

Extensions to the STAR syntax will not, and should not, happen frequently, but when significant gains can be made to a data model with new language extensions, they should, and will, occur.

5.3 Why is a record of data origin important?

The importance of knowing the *origin,* or source, of data items is often undervalued in data management practices. Data items in any domain may be divided into two basic origin-types: *primitive* and *derivative*. Further divisions can be made but for most definition purposes these two are sufficient. Primitive data items cannot be derived from data relationships with other items; whereas derivative items can. The values of primitive items may be determined by observation (*e.g.* shape), measurement (*e.g.* length) or postulation. Derivative item values can be evaluated from primitive or other derivative items using *prior knowledge*. For example, density is usually deemed a derivative item because its value can be calculated from the values of mass and volume items; which themselves can be either primitive or derivative items. The point to be made here is that the value of a derivative item is dependent on the current knowledge at the time of its derivation. This means that its value is time-dependent. Primitive item values are not, though it is possible that their precision may improve with time because of advances in instrumentation and techniques.

Is there any difference in the relative archival importance of data in each origin-type? Primitive data, at the very least, must be of equivalent archival value to derivative data! Past publication and database practices often ignored primitive data and retained only information that could easily be derived. Fortunately such practices no longer exist in most disciplines, and raw data is now routinely archived and will be available to future methodologies. For older published work, derived results can only be recalculated after the raw data is recollected.

The time-variability of data is a powerful reason why data ontologies are essential for recording the precise nature of items and their relationships. Time-stamped ontologies record the evolution of methodologies. In the future it will be just as important to have the domain knowledge of 50 years ago, as it is to have today's. The STAR approach, with a carefully crafted DDL ontology, allows both the origin of a data item to be made explicit, and time-stamping of definitions with evolving interpretations to be implemented.

5.4 Why should domain ontologies be shared?

The fundamental premise of any tag–value approach to information retention is that every tag in a domain is absolutely unique. It is the unique key that opens the appropriate ontology definition. One may readily argue that, as with website addresses, tags should also be unique *across* domains. It is also apparent that many data items have identical equivalences across domains and sub-domains. In order not to duplicate the metadata of these items, their definitions need to be shared, otherwise ontology maintenance becomes too complicated. For this reason the ontology protocol DDLm (Spadaccini & Hall 2012b) allows definition material in one ontology to be included in another. This may also be used to simplify the human-readability of ontologies by 'hiding' repetitive definition components, and large tables such as enumeration states, in an auxiliary file. These definition components are automatically imported during the machine parsing of an ontology.

5.5 Why is exact data relationships information important to ontologies?

Wherever possible precise relationships between data items should be recorded in the method attribute. In STAR ontologies, the method record is a mathematical expression written in the symbolic script language *dREL* (Spadaccini, Castleden, du Boulay & Hall 2012). Methods represent essential knowledge on the nature

of a derivative item, and allow, for each data instance, their evaluation and validation from related item values. It should be stressed that method scripts are not intended to duplicate or replace computer programs dedicated to these calculations. The method expression is simply a *complete description*, exact because it is recorded in mathematics, of a data item in terms of other items. Although some ontology tools, such as *JsCif* (du Boulay 2013), can execute these methods for a specific data instance, they are intended only as a benchmark for evaluating data against their definition, or for validating values in a data instance with these evaluations. An ontology will probably never ever match the speed and flexibility of software dedicated to large scale computation, but it will provide an important gold standard for checking the correctness of such software.

6 Concluding remarks

In this paper, we have made the case for ontologies as drivers for effective communication and validation of structured information. We have given a lengthy account of the implementation of such an information system in the specific field of crystallography, in part to demonstrate the potential benefits, and in part to encourage other disciplines that are not building ontology-based information systems to do so. Our experience demonstrates that there are technical challenges in setting up an ontology-based workflow. These include choice of an appropriate UDL, development of new software (or modification of existing libraries), and the often underestimated labour of codifying the relevant concepts within a discipline.

However, these are tractable, especially when there are existing systems that can be used as models – we are happy to offer the STAR approach as one possibility for communities that are not already invested in a different one. At the technical level, a relatively small team of committed developers can build the ontologies themselves and the necessary supporting software infrastructure.

The major hurdles, however, are sociological. We have emphasized the important role of the IUCr in crystallography as a respected and authoritative sponsor and enabler. Where a scholarly community cannot call upon such an authority for support, it must actively work to educate its users in the potential benefits of adopting this new approach to collecting, analyzing and exchanging scientific data and other structured information. One particular advantage of extensible ontologies, that might assist in what is often an uphill task, is that the endeavour can begin with relatively small, well-focused areas of study, and progressively expand as the first fruits of the invested effort become recognized.

7 Competing Interests

The authors declare that they have no competing interests.

8 Acknowledgements

We are grateful to the International Union of Crystallography for its encouragement and support of the Crystallographic Information Framework project over almost three decades. SRH acknowledges the practical support of the University of Western Australia. We are indebted to the dozens of far-sighted colleagues, many of them recognized in the reference list, who have worked to make the STAR and CIF approaches so effective.

9 References

Allen, F. H., Barnard, J. M., Cook, A. F. P., & Hall, S. R. (1995) The Molecular Information File (MIF): Core Specifications of a New Standard Format for Chemical Data. *J Chem. Inf. Comput. Sci., 35*: 412–427. DOI: http://dx.doi.org/10.1021/ci00025a009

Allen, F. H., Johnson, O., Shields, G. P., Smith, B. R., & Towler, M. (2004) CIF applications. XV. *enCIFer*: a program for viewing, editing and visualizing CIFs. *J. Appl. Crystallogr., 37*: 335–338. DOI: http://dx.doi.org/10.1107/S0021889804003528

Bernstein, H. J., & Hammersley, A. P. (2005) Specification of the Crystallographic Binary File (CBF/imgCIF). Chap. 2.3 of *International Tables for Crystallography,* Vol. G. First edition. Dordrecht: Springer.

Biological Magnetic Resonance Data Bank. Available at: www.bmrb.wisc.edu.

Brown, I. D. (1988) Standard Crystallographic File Structure-87. *Acta Crystallogr. A44*: 232. DOI: http://dx.doi.org/10.1107/S010876738700970X

Brown, I. D., & McMahon, B. (2006) The Crystallographic Information File (CIF). *Data Science J., 5*: 174–177. DOI: http://dx.doi.org/10.2481/dsj.5.174

Davies, A. N. (2002) *Chemistry International*, *24*, 3–8. Available at: http://www.iupac.org/publications/ci/2002/2404/index.html.

du Boulay, D (2013) JsCif: JavaScript for displaying ontology definitions and uploaded instance data with evaluation and validation. Available at: www.iucr.org/resources/cif/software.

Florabase – The Western Australian Flora. Available at: www.florabase.dpaw.wa.gov.au.

Gruber, T. R. (1993) A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, *5*: 199–220. DOI: http://dx.doi.org/10.1006/knac.1993.1008

Hall, S. R. (1991) The STAR File: A New Format for Electronic Data Transfer and Archiving. *J. Chem. Inf. Comput. Sci., 31*: 326–333. DOI: http://dx.doi.org/10.1021/ci00002a020

Hall, S. R. (1993a) CIF Applications II. CIFIO: for CIF Input/Output in the Xtal System. *J. Appl. Crystallogr., 26*: 474–479. DOI: http://dx.doi.org/10.1107/S0021889893000871

Hall, S. R. (1993b) CIF Applications III. CYCLOPS: for validating CIF Data Names. *J. Appl. Crystallogr., 26*: 480–481. DOI: http://dx.doi.org/10.1107/S0021889893000883

Hall, S. R. (1993c) CIF Applications IV. CIFtbx: a toolbox for manipulating CIFs. *J. Appl. Crystallogr., 26*: 482–494. DOI: http://dx.doi.org/10.1107/S0021889893050897

Hall, S. R., Allen, F. H., & Brown, I. D. (1991) The Crystallographic Information File (CIF): a new standard archive file for crystallography. *Acta Crystallogr., A47*: 655–685. DOI: http://dx.doi.org/10.1107/S010876739101067X

Hall, S. R., & McMahon, B. (eds.) (2005) Definition and Exchange of Crystallographic Data. *International Tables for Crystallography,* Vol. G. First edition. Dordrecht: Springer.

Hall, S. R., & Sievers, R. (1993) CIF Applications I. QUASAR: for extracting CIF data. *J. Appl. Crystallogr., 26*: 469–473. DOI: http://dx.doi.org/10.1107/S002188989300086X

Hester, J. (2015) Creating and manipulating universal metadata definitions. Workshop on Metadata for raw data from X-ray diffraction and other structural techniques, Rovinj, Croatia, 22–23 August 2015. Available at: http://www.iucr.org/resources/data/dddwg/rovinj-workshop and https://youtu.be/gMCK5pbHI9o.

Murray-Rust, P. (2014) *Personal communication by email.*

Murray-Rust, P., & Rzepa, H. S. (1999) Chemical Markup, XML, and the Worldwide Web. 1. Basic Principles. *J. Chem. Inf. Comput. Sci., 39*: 928–942. DOI: http://dx.doi.org/10.1021/ci990052b

Spadaccini, N., Castleden, I. R., du Boulay, D., & Hall, S. R. (2012) dREL Relational Expression Language for Dictionary Methods. *J. Chem. Inf. Model., 52*, 1917–1925. DOI: http://dx.doi.org/10.1021/ci300076w

Spadaccini, N., & Hall, S. R. (2012a) Extensions to the STAR File Syntax. *J. Chem. Inf. Model., 52*: 1901–1906. DOI: http://dx.doi.org/10.1021/ci300074v

Spadaccini, N., & Hall, S. R. (2012b) DDLm: A New Dictionary Definition Language. *J. Chem. Inf. Model., 52*: 1907–1916. DOI: http://dx.doi.org/10.1021/ci300075z

Spadaccini, N., Hall, S. R., & McMahon, B. (2005) STAR File Utilities. Chap. 5.2 of *International Tables for Crystallography,* Vol. G. First edition. Dordrecht: Springer.

Spadaccini, N., Mildenhall, G., & Hall, S. R. (2008) Star <->XML tools: links to W3C exchange standards. *Acta Crystallogr., A58*(Supplement): C257. DOI: http://dx.doi.org/10.1107/S0108767302095272

Thompson, H. S., Beech, D., Maloney, M., & Mendelsohn, N. (2004) XML Schema Part 1: Structures Second Edition. Available at: http://www.w3.org/TR/xmlschema-1/#normative-schemaSchema.

Westbrook, J. D., Henrick, K., Ulrich, E. L., & Berman, H. M. (2005) The Protein Data Bank exchange data dictionary. App. 3.6.2 of *International Tables for Crystallography,* Vol. G. First edition. Dordrecht: Springer.

Westrip, S. P. (2010) *publCIF*: software for editing, validating and formatting crystallographic information files. *J. Appl. Crystallogr., 43*: 920–925. DOI: http://dx.doi.org/10.1107/S0021889810022120

How to cite this article: Hall, S R and McMahon, B 2016 The Implementation and Evolution of STAR/CIF Ontologies: Interoperability and Preservation of Structured Data. *Data Science Journal,* 15: 3, pp. 1-15, DOI: http://dx.doi.org/10.5334/dsj-2016-003

Submitted: 11 June 2015 **Accepted:** 25 November 2015 **Published:** 12 January 2016

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS