Conference paper

David Rauh*, Claudia Blankenburg, Tillmann G. Fischer, Nicole Jung, Stefan Kuhn, Ulrich Schatzschneider, Tobias Schulze and Steffen Neumann

Data format standards in analytical chemistry

https://doi.org/10.1515/pac-2021-3101

Abstract: Research data is an essential part of research and almost every publication in chemistry. The data itself can be valuable for reuse if sustainably deposited, annotated and archived. Thus, it is important to publish data following the FAIR principles, to make it findable, accessible, interoperable and reusable not only for humans but also in machine-readable form. This also improves transparency and reproducibility of research findings and fosters analytical work with scientific data to generate new insights, being only accessible with manifold and diverse datasets. Research data requires complete and informative metadata and use of open data formats to obtain interoperable data. Generic data formats like AnIML and JCAMP-DX have been used for many applications. Special formats for some analytical methods are already accepted, like mzML for mass spectrometry or nmrML and NMReDATA for NMR spectroscopy data. Other methods still lack common standards for data. Only a joint effort of chemists, instrument and software vendors, publishers and infrastructure maintainers can make sure that the analytical data will be of value in the future. In this review, we describe existing data formats in analytical chemistry and introduce guidelines for the development and use of standardized and open data formats.

Keywords: Analytical chemistry; cheminformatics: data and standards; data standard; file format; mass spectrometry; NMR.

Introduction

The amount of research data is growing in all disciplines, and chemistry is no exception. Modern methods and instruments effortlessly produce large amounts of data, stored within the data infrastructure of research institutes. Some of that data might be published supplementing scientific articles, but the majority of data accessible currently does not lead to facile reuse of the data by other scientists or is readily available for computational analysis such as machine learning. The reasons for this situation are manifold: Even if most of the journal articles are published in digital form, important information is often hidden in text and images or even left out entirely for publication. If accessible, the required information is frequently given as free-form textual description [1]. Tables or schematic representations of *e.g.* chemical reactions in the article PDF files are

Article note: A collection of invited papers on Cheminformatics: Data and Standards.

***Corresponding author: David Rauh**, Leibniz Institute of Plant Biochemistry, Bioinformatics and Scientific Data, Weinberg 3, 06120 Halle, Germany, e-mail: drauh@ipb-halle.de. https://orcid.org/0000-0001-7499-1693
Claudia Blankenburg, Tillmann G. Fischer and Steffen Neumann, Leibniz Institute of Plant Biochemistry, Bioinformatics and Scientific Data, Weinberg 3, 06120 Halle, Germany, e-mail: cblanken@ipb-halle.de (C. Blankenburg), tfischer@ipb-halle.de (T.G. Fischer), sneumann@ipb-halle.de (S. Neumann). https://orcid.org/0000-0002-7899-7192 (S. Neumann)
Nicole Jung, Karlsruhe Institute of Technology, Institute for Chemical and Biological Systems (IBCS-FMS), Hermann von Helmholtz Platz 1, 76344 Eggenstein-Leopolshafen, Germany, e-mail: nicole.jung@kit.edu
Stefan Kuhn, School of Computer Science and Informatics, De Montfort University, Leicester, UK, e-mail: stefan.kuhn@dmu.ac.uk
Ulrich Schatzschneider, Institut für Anorganische Chemie, Julius-Maximilians-Universität Würzburg, Am Hubland, D-97074 Würzburg, Germany, e-mail: ulrich.schatzschneider@uni-wuerzburg.de
Tobias Schulze, Department of Effect-Directed Analysis, Helmholtz Centre for Environmental Research – UFZ, Permoserstr. 15, 04318 Leipzig, Germany, e-mail: tobias.schulze@ufz.de

often challenging to convert back to tabular formats for reuse. Their contextualization can be difficult due to missing common standards for metadata. The situation is even worse regarding data files of instrumental measurements, which are often recorded in proprietary vendor formats and impossible to be opened without the vendors' software.

To overcome this situation, digital assets can (and should) be shared based on the FAIR principles for scientific data management [2]. These principles define that digital objects should be findable, accessible, interoperable and reusable to be of value for (different) potential consumers – human beings or computers. Findability and accessibility can be achieved by depositing data in public repositories and databases, which also ensures that all data is accompanied by a defined set of metadata. Interoperability and reusability depend on data standards to describe which data and metadata should be recorded and how they are made available. In other words, standards are prerequisites of FAIR data. Other studies already underlined the importance of data standards to prevent information decay [3], curation of old datasets with new methods [4] or the reuse of experimental raw data [5, 6]. Some publishers also started to encourage authors to submit data following the FAIR principles.

The availability and use of open standards to obtain FAIR data is key (Fig. 1). Thus, it is necessary to recommend and apply existing open standards or, if they do not suffice, to extend them or create new open standards for archival, exchange and reuse of experimental data and metadata. Standardized open formats allow reading, analyzing, and reusing data collected from different sources and enable integrated, comprehensible and reusable workflows. Lowering the technical barriers to access shared data will improve the quality of data, by simplifying the processes of review, quality control, comparison and reproduction [7]. Not

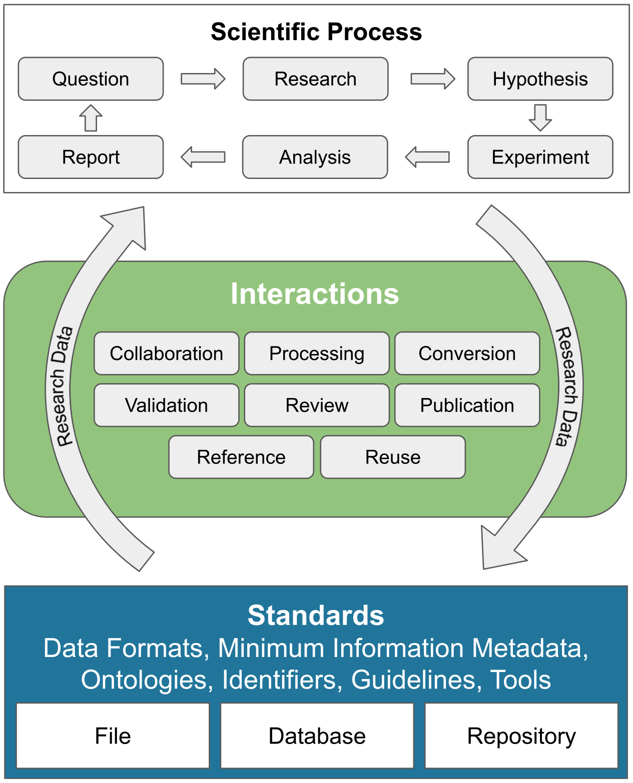


Fig. 1: The role of data standards to share research data. Any interaction coupled to sharing research data via data files, databases or repositories requires standards to define data formats, minimal information metadata, ontologies, identifiers, guidelines and tools. These facilitate findability in conjunction with persistent identifiers (PIDs) as well as accessibility and interoperability for humans and computers to enable reuse.

only the quality of data itself, but also the quality of software applied in digital workflows can be enhanced, if software developers can rely on standards when implementing and maintaining servers, databases or tools.

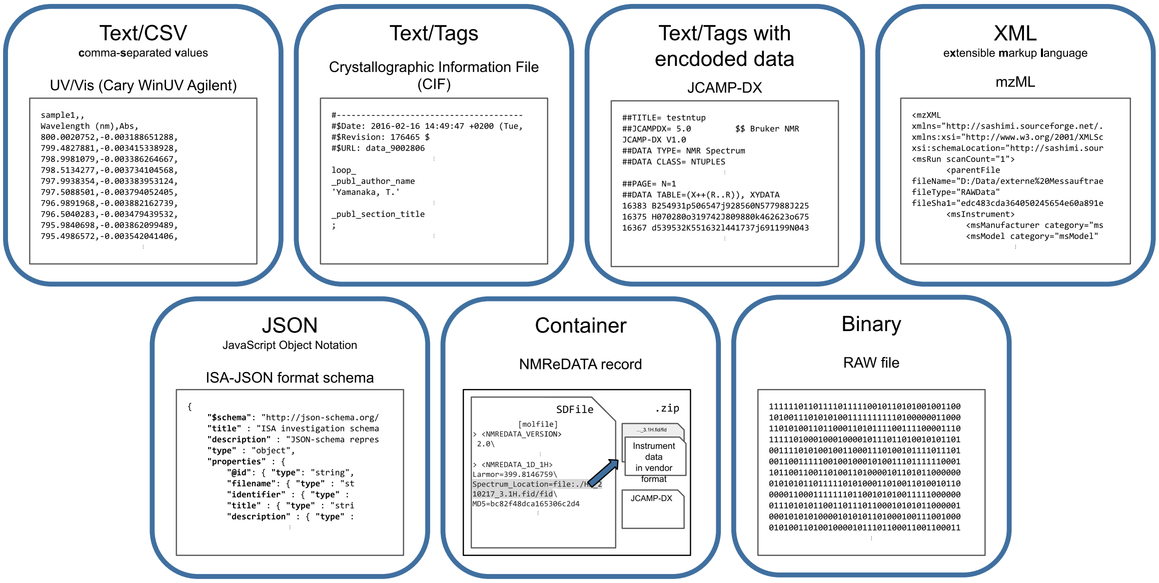
In this article, we will review the state of data standards and file formats in chemistry for selected analytical techniques including mass spectrometry, nuclear magnetic resonance spectroscopy, infrared and UV/Vis spectroscopy as well as X-ray diffraction and X-ray absorption and emission spectroscopy. Additionally, we summarize standardization efforts and bodies in chemistry and start with definitions of terms, which chemists might not be familiar with, but which are frequently used in this review.

Relationship between data models, specifications, file formats and software implementations

For successful data standards, several components are required: the data model and the data representation. A specification of a representation may contain the model implicitly.

The **Data model** describes how data is organized, which information they contain, the data types (*e.g.* text, numbers, lists), the relationship between these components, and rules of component and data integrity. Conceptually, the model and its components can be organized in different ways, *e.g.* flat, multidimensional, as network or hierarchical. The meaning of the applied components has to be well described in an unambiguous way. Ontologies can be consulted to ensure a consistent usage of the components in the context of the model. As the data model is abstract, it is necessary to implement **representations** for it, such as a database or data file on a computer.

File formats are one way to represent a data model. File formats as models can be categorized by different criteria: proprietarily vs. openly specified, binary vs. text, simple vs. complex or flat vs. n-dimensional (Fig. 2). For long-term storage, a compact binary format may be the first choice, while for further processing with cheminformatics tools a format based on standards like Comma Separated Values (CSV), Extensible Markup Language (XML) or JavaScript Object Notation (JSON) may be more favourable due to support by most programming languages. When thinking about data strategies, open formats should be considered. Proprietary


Text/CSV
comma-separated values
UV/Vis (Cary WinUV Agilent)

```
sample1,,
Wavelength (nm),Abs,
800.0020752,-0.003188651288,
799.4827881,-0.003415338928,
798.9981079,-0.003386264667,
798.5134277,-0.003734104568,
797.9938354,-0.003383953124,
797.5088501,-0.003794052405,
796.9891968,-0.003882162739,
796.5040283,-0.003479439532,
795.9840698,-0.003862099489,
795.4986572,-0.003542041406,
              ⋮
```

Text/Tags
Crystallographic Information File (CIF)

```
#-------------------------------------
#$Date: 2016-02-16 14:49:47 +0200 (Tue,
#$Revision: 176465 $
#$URL: data_9002806
              ⋮
loop_
_publ_author_name
'Yamanaka, T.'
              ⋮
_publ_section_title
;
              ⋮
```

```
##TITLE= testntup
##JCAMPDX= 5.0         $$ Bruker NMR
JCAMP-DX V1.0
##DATA TYPE= NMR Spectrum
##DATA CLASS= NTUPLES
              ⋮
##PAGE= N=1
##DATA TABLE=(X++(R..R)), XYDATA
16383 B254931p506547j928560N577988J225
16375 H070280o319742J809880k462623o675
16367 d539532K551632l441737j691199N043
              ⋮
```

```
<mzXML
xmlns="http://sashimi.sourceforge.net/.
xmlns:xsi="http://www.w3.org/2001/XMLSc
xsi:schemaLocation="http://sashimi.sour
<msRun scanCount="1">
       <parentFile
fileName="D:/Data/externe%20Messauftrae
fileType="RAWData"
fileSha1="edc483cda364050245654e60a891e
       <msInstrument>
           <msManufacturer category="ms
           <msModel category="msModel"
              ⋮
```

```
{
    "$schema": "http://json-schema.org/
    "title" : "ISA investigation schema
    "description" : "JSON-schema repres
    "type" : "object",
    "properties" : {
        "@id": { "type": "string",
        "filename": { "type" : "st
        "identifier" : { "type" :
        "title" : { "type" : "stri
        "description" : { "type" :
              ⋮
```

```
        [molfile]
> <NMREDATA_VERSION>
 2.0\
            ⋮
> <NMREDATA_1D_1H>
Larmor=399.8146759\
Spectrum_Location=file:./H_2
10217_3.1H.fid/fid\
MD5=bc82f48dca165306c2d4
            ⋮
```

```
1111110110111101111100101101010010011OO
```

Fig. 2: Examples for different data file formats. XML and JSON are similar in their logical structure, but use different syntax to encode the data model. Binary files may also have an internal structure, but reading and access requires dedicated software libraries.

formats may prevent further reuse due to issues connected to licensing, poor documentation and support by the vendor. Complex data often need more complex formats to express the meaning and relation of data.

To represent data distributed over various files or to bundle files from different stages of the workflow, container file formats like ZIP can maintain the folder structure and provide compression. Index files or cross references may link these files in the container context.

The **specification** of a data file standard may describe the whole format from scratch or can utilize generic formats, which are standardized anywhere else. Such generic formats include CSV [8], XML [9] or JSON [10]. For some of these generic formats, a formal specification format is available, called the schema format, which streamlines the specification of format derivatives. XML or JSON files can be described by schemas. If the format cannot be described formally by a general or existing schema, it is important to create an unambiguous new specification, clear in description of syntax, and grammar, as well as in definition of the components of the underlying data model. File formats should also be extendable to allow the usage of the format for new techniques and with future requirements. Data file standards not fulfilling these criteria may fork into incompatible formats or will be outdated soon and replaced by new standards.

Any software importing or exporting a file format has to include readers and/or writers for this format, often accompanied by a validator for checking documents for adherence to the standard. Integration of any format into software is more likely if developers may rely on an **implementation**, available as a library, package or module for their programming language or environment. Thus, acceptance of a format also depends on the availability and support of implementations. Formats are more sustainable if they have a free licence and are open source. For formats described by schemas there are often well-maintained implementations for the generic format including validation. So only the schema file has to be provided. Implementations for formats, which are not openly specified and lack vendor support, are sometimes only feasible by reverse engineering, which is debatable concerning legality. Only a complete and unambiguous specification allows well-crafted implementations.

In this paper, we do not deal with the question of how to package and store data. This is a separate field, which has recently been addressed *e.g.* by RO-Crate [11], which describes ways how to link data stored in diverse locations.

Standardization efforts and -bodies in chemistry

To be adopted by the community of scientists, any data format or model is required to be well-defined and to include necessary and valuable information. This applies to standards developed under the auspice of an official standardization body and to community-driven efforts. Nonetheless, standardization bodies might guarantee long-term maintenance as well as further developments to adapt to new techniques and demands. Data standards can be created by community initiatives or through official generic or discipline focused standardization bodies. Some of these organizations and initiatives are listed in Table 1.

The **International Union of Pure and Applied Chemistry** (IUPAC) with its numerous working groups and committees [12] is a major player in the field of standardization in chemistry. Among other topics, the Subcommittee on Cheminformatics Data Standards is the maintainer of the JCAMP-DX format standard [13].

The **Allotrope Foundation** is "a group of pharmaceutical, device vendor, and software companies that develops and releases technologies […] to simplify the exchange of electronic data." [14], established nearly 10 years ago. The foundation is working on 1) the Allotrope Data Format (ADF) which is based on the widely used domain independent hierarchical data format 5 (HDF5) combined with semantic metadata described via the resource description framework (RDF), 2) Allotrope Data Models (ADM), and 3) Allotrope Foundation Ontologies (AFO). Only the latter is licensed under Creative Commons terms. ADF is used in several industrial environments, since several instrument providers are members of the foundation and provide Allotrope compatible software interfaces for their devices. With the Allotrope Simple Model (ASM) there is a way to make data more accessible by using the popular JSON format.

Table 1: Selected organizations, consortia and initiatives developing and promoting standards and open data for chemistry and other fields.

Short name	Full name	Organizational form	Main focus	References
Allotrope	Allotrope Foundation	Consortium	Data lifecycle	[14]
CODATA	Committee On Data	Committee	Open science	[19]
DataCite	DataCite	DOI registration agency	DOIs	[23]
fairsharing.org	FAIRsharing	Registry	Policies, databases, standards	[25]
FORCE11	The Future of Research Communication and e-Scholarship	Community	Publication standards	[21]
GO FAIR	GO FAIR Initiative	Initiative	FAIR data	[18]
IUPAC	International Union of Pure and Applied Chemistry	Organization	Standards for chemistry	[12]
NFDI4Chem	Nationale Forschungsdaten-Infrastruktur für die Chemie [National research data infrastructure for chemistry]	Consortium	Infrastructure, standards	[17]
RDA	Research Data Alliance	Initiative	Open data	[20]
re3data.org	Registry of Research Data Repositories	Registry	Repositories	[26]

The **Pistoia Alliance** is also a non-profit organization of more than 100 member companies, working in several subprojects. One of those is the Unified Data Model (UDM) [15] which covers experimental information on the synthesis of compounds and their testing, including referencing of analytical data, and the possibility to use the AFO. The UDM licence was recently changed to the liberal MIT licence, which should facilitate the wider adoption of this model.

The consortium for **Standardization in Lab Automation** (SiLA) is a non-profit initiative of software and device suppliers to standardize software interfaces for laboratory automation. Alongside their communication protocol for instruments, they promote the AnIML format (described below) for data storage and interchange [16].

NFDI4Chem is a national consortium in Germany [17], embedded in the interdisciplinary NFDI (Nationale Forschungsdateninfrastruktur [National Research Data Infrastructure]), which will contribute in developing and specifying recommendations for analytical data standards in chemistry for users but also for the instrument and software providers, as well as providing metadata and publication standards.

There are several other organizations to support the science community to make their data more FAIR and more open. **GO FAIR** is a worldwide acting initiative supporting the community in establishing FAIR principles by raising awareness, providing training and giving recommendations for technical standards and infrastructure [18]. The **Committee on Data** (CODATA) promotes Open Science [19], and thus, also FAIRness of data and open standards as requirements. The community-driven **Research Data Alliance** (RDA) has a similar mission of encouraging scientists to share and reuse their data [20].

There is currently no general agreement by the publishers on how to publish and cite data sets. Communities such as **FORCE11**, with currently about 3500 members [21], support initiatives in harmonizing publication standards including data citation. One of the prerequisites to make data citable is the existence of persistent identifiers (PIDs) [22]. The most widely adopted PIDs are Digital Object Identifiers (DOIs), with CrossRef as the main DOI registration agency for publications and DataCite as a provider of DOIs for datasets [23, 24].

Although these initiatives can help to raise awareness for the challenges connected to RDM and provide strategic advice in improving RDM, chemists need domain-specific information on existing standards, data models and file formats. Moreover, resources on repositories for datasets or databases with domain-specific information are required. NFDI4Chem will provide recommendations to this domain specific needs and will contribute to standard development for data exchange and archival by using, extending and creating open data formats.

There are some registries present which can be used as a starting point. **FAIRsharing** provides information on policies, standards and databases for many subjects including chemistry [25]. The **Registry of Research Data Repositories** (Re3data) also offers searchable information for domain-specific and discipline-agnostic repositories and databases [26].

Data formats for analytical data

When developing infrastructures for chemistry, one of the major tasks is to adopt or define data models and formats for exchange, storage and archival of analytical data. Vendors often provide proprietary formats to store instrumental and experimental metadata as well as data readings. With the growing set of workflows, software tools, databases and repositories there is demand for open and well specified models and format standards. It has also become evident that broad adoption of a standard data format benefits from reference software implementations, *i.e.* software modules for common programming languages with the main purpose to read and write a specific format, that can easily be reused in larger software projects.

A simple practice to export and exchange analytical data is conversion to text based and tabular formats, like CSV. These are human-readable and can usually be interpreted by the user. However, with the lack of metadata and specification, these formats are less appropriate for automatic processing and storage. Nevertheless, implementation of text based and tabular file format support is straightforward. Therefore, there are several approaches to create extensible formats to store an unified set of metadata together with experiment-specific metadata and experimental results. Most of them are joint efforts of scientific institutions and/or researching companies (Table 2).

JCAMP-DX is a format which can be applied for a wide range of analytical data. It was developed by the Joint Committee on Atomic and Molecular Physical Data (JCAMP) since 1988 [27] and is now maintained under the auspices of IUPAC. A general standard is proposed which can be used for different spectroscopic and spectrometric methods. Additionally, defined special standards for electron paramagnetic resonance (EPR) [13] and nuclear magnetic resonance (NMR) [28], chromatography and mass spectrometry [29] were published. Because the standard does not provide native support for ontology or controlled vocabulary use, and each implementation may use its own extensions, files from different sources can be incompatible. There is a Java

Table 2: Some common file formats for analytical data.

Format	Data type	Maintainer	Parent format	Specification	References
JCAMP-DX	Multiple	IUPAC	ASCII, text	Open	[14, 27–29]
AnIML	Multiple	ASTM	XML	Open	[34, 35]
netCDF	Multiple	UCAR	CDF	Open	[31]
CSV	Multiple	IETF-RFC	ASCII, text	Open	[8]
Text, ASCII	Multiple	(Open)		Self explanatory	
ISA	Multiple	ISA Commons Community	TSV or JSON	Open	[33, 36]
UDM	Multiple	Pistoia Alliance	XML	Open	[15]
ADF	Multiple	Allotrope	HDF5+RDF	For members	[14]
mzML	Mass spectrometry	HUPO/PSI	XML	Open	[37, 38]
ANDI-MS	Mass spectrometry	ASTM International	netCDF	Open	[39]
nmrML	NMR	COSMOS	XML	Open	[40]
NMReDATA	NMR	NMReDATA Initiative	SDF	Open	[41, 42]
CIF	X-ray diffraction	IUCr	Text	Open	[43, 44]
Bruker FID	NMR	Bruker	(Binary)	Proprietary	
mnova	NMR	Mestrelab	(Binary)	Proprietary	
Bruker OPUS	Spectroscopy	Bruker	(Binary)	Proprietary	
Perkin Elmer	Spectroscopy	Perkin Elmer	ASCII, text	Proprietary	
ThermofFisher Grams	Spectroscopy	ThermoFisher	Binary	Proprietary	

reference implementation for this format available and there are libraries for other programming languages applicable like Python, R JavaScript and MATLAB. As JCAMP-DX is accepted as the exchange format for many analytical methods, there is a wide support by software for spectral analytics.

The XML-based **Analytical Information Markup Language (AnIML)** has been created to be an ASTM International standard and covers different analytical techniques. The standard comprises schema definitions for a generic core and technique-specific documents. Thus, it is possible to define technique documents for various analytical measurements. As AnIML is fully specified by its XML schemas, it can be effortlessly implemented in any language with XML support. No reference implementation is available, but among the (few) open-source implementations Jmol/JSmol (formerly JSpecView) can import and visualize AniML documents. For developers working with the python programming language, a library is under development to create, parse and validate AniML files [30]. There is also support by BSSN Software (now Merck) that promotes the format, often in combination with the device interface SiLA (Standardization in Lab Automation).

NetCDF is a binary file format and software interface, mainly defined by its implementations by the Unidata community [31]. It is an abstract model, which can be extended by self-describing objects. Thus, this model can be flexibly adopted to specific use cases. A family of ANDI (ANalytical Data Interchange) formats is specified by the ASTM, which are based on netCDF (see also ANDI-MS below).

The **ISA (Investigation-Study-Assay)** framework, originated from the bioscience community, defines the hierarchical ISA data model to store metadata on project context and study details, and analytical measurement data [32]. As the abstract ISA data model already encourages the user to annotate any parameter or value with ontology terms, it assures well described datasets. Implementations are available as tab separated value files (ISA-Tab) or as JSON (ISA-JSON). The ISA API is a Python library implementing the model for usage with the ISA formats [33]. The model is also applied for repositories such as MetaboLights. Moreover, journals such as ScientificData or GigaScience use the ISA data model to describe complex experimental setups covered in the manuscripts.

Data standards in UV/Vis-, IR- and Raman-spectroscopy

Experimental data obtained with spectroscopic methods such as infrared spectroscopy, Raman and UV/Vis-spectroscopy, are often comparatively small in size and straightforward in their structure. Vendors store the raw data in proprietary formats, either as binary data or in ASCII. These can be exported as (or converted to) Excel spreadsheets or plain text tables with x,y pairs (CSV or similar format). A header section may include metadata.

Users have to further process such data for their specific needs, and currently no overarching specifications exist. Repositories may convert these to a specified format, *e.g.* Chemotion ELN will convert text and excel files to JCAMP-DX.

There are a few vendor formats which are popular for data exchange: GRAMS SPC, Perkin Elmer SP and Bruker OPUS files are supported not only by the format creators, but also by other vendors and instrument-agnostic software tools for *e.g.* statistical analysis.

There were efforts to create a special format for ultraviolet-visible spectroscopy data, which was called SpectroML [45, 46], which have now been superseded by the more general AnIML [34, 35]. Harmonization between instrument vendors and adoption of an open standard still needs to be achieved.

Data standards for NMR spectroscopy

NMR is an indispensable analytical technique providing rich information on bonding and structure as well as molecule interaction and abundance of molecules in samples. Until now, it was common practice to publish the spectra as images in supplementary materials, regularly published as PDF files. Additionally, a list of shifts is reported, sometimes referred to as NMR text.

However, the raw data containing the Free Induction Decays (FIDs), initially processed spectra and the instrument metadata is usually not published, which might allow for reanalysis and reuse. The importance of providing both FID raw data and extracted NMR spectra were previously demonstrated extensively [5, 47].

All instrument vendors have developed their own (binary) raw data formats. Since this FID data itself is mainly time-response data with a straight-forward structure, most of the vendor formats are supported by the software used within the NMR community. Many vendors also agreed to import and export the JCAMP-DX format, which has a specification for FID raw data and is recommended by IUPAC [28, 48].

The JCAMP-DX format can also be used for exchange, import and export of multidimensional spectra. Because of the open and extensible nature of JCAMP-DX on the one hand and the lack of a controlled vocabulary on the other, there are already different flavours of the NMR format implementation, hence, validation or import might be challenging.

Inspired by the mass spectrometry format mzML (see below) the standard **nmrML** [49] was initially developed for metabolomics data, but can also be used for any other kind of NMR data obtained. The standard nmrML is a XML based format for FID raw data for 1D as well as 2D NMR spectra. Due to the explicit syntax specification of this format and the underlying controlled vocabulary (nmrCV), data files can be validated. It is used as a storage format for NMR data in the Metabolights data repository [50].

The open **NMReData** format is maintained by the NMReData initiative [41, 42]. The NMR record in NMReDATA format includes the instrument (raw) data, a SDFile and, since version 2, also spectral data in JCAMP format in a folder, which can be compressed in the zip format for data exchange. The SDFile contains the chemical structure and the actual NMReDATA as standardized SDF tags. These tags take account of chemical shifts, couplings, signal assignments and lists of 2D correlations, only to mention a few. NMReData can be used for 1D and 2D spectra and contains a core set of NMR parameters. The format allows raw data, extracted data, and structures to be recorded in one format, which is not fully supported in existing formats. It is machine- and human-readable at the same time, and allows flexibility and extensions. For exchange a NMReData record can be compressed in ZIP format. FAIRness is the overall principle behind it. Members of the NMReData initiative include open-source projects such as NMRShiftDB2 and Cheminfo.org, commercial NMR software vendors such as MestreLab, NOMAD, C6H6 and ACD/Labs and device vendors like Bruker [41, 42].

Recently, there were several additional open standard formats developed by the NMR community. A great effort came from protein structure determination by NMR, a field where the specifics and size of the macromolecules required special data formats [51].

Derived from the self-defining STAR format are the also more protein specific **NMR-STAR** [52] which is used by the BioMagResBank (BMRB) data format and defines over 4600 data item tags describing data and metadata, which are organized in more than 300 categories and 80 category groups. The **NMR Exchange Format (NEF)** [53] format was developed for storage of NMR data in wwPDB. It is more accessible for software developers by reducing the complexity. Additionally, it is extensible with application-specific tags. As NMR-STAR and NEF both are derived from the STAR format, they are convertible, and the only formats accepted by wwPDB and BMRB. The Collaborative Computing Project for NMR (CCPN) is developing NEF [51], based on the data model [53] for usage within their protein NMR focused software tools.

Data standards in mass spectrometry

A distinction that is rather important in different disciplines of chemistry is whether a particular spectrum is the data of interest, or whether a set of spectra shall be represented. In the former case, text-based file formats like JCAMP-DX [29], Mascot Generic File (MGF) [54] or National Institute for Standards and Technology Mass spectrometry (NIST MSP) [55] may be sufficient. However, for entire runs using, *e.g.*, LC-MS or GC-MS with hundreds of chromatography-resolved spectra, more efficient file formats have been developed. The netCDF (Network Common Data Form) based **Analytical Data Interchange Protocol for Mass Spectrometry (ANDI-MS)** is an ASTM International standard [39]. It was developed initially as an Analytical Instrument Association (AIA) standard as a follow-up of the ANDI for Chromatographic Data specification. Technically, it builds upon NetCDF [31], a generic and highly efficient container format. The ANDI-MS specification defines which elements are needed to encode mass spectrometry data.

More complex MS experiments require capturing a rich set of instrumental settings such as per-scan polarity, isolation windows and collision energies. Several formats (mzXML, mzData) had been developed in the early days of proteomics [56, 57], which have been merged into **mzML** by the Proteomics Standards Initiative (PSI) [38, 58]. Despite the term Proteomics in its name, many of the PSI standards can also be used for respective analytical data from samples beyond Proteomics. The XML based mzML data format is a widely accepted standard for analytical mass spectrometry data, recommended by several societies and infrastructures for data exchange and archival. There is also a wide range of tools, including converters and spectra viewers, and software libraries to work with mzML files [33, 47–52]. The use of the PSI-MS ontology as controlled vocabulary, combined with data validators, provides excellent interoperability between consumers and producers of mzML, regardless of the instrument vendor or analysis software.

The XML-based nature of these formats ensures that the data is readable by most, if not all, computer systems and programming languages long-term. To improve performance for fast random access and parallel processing of data, the same data model was used in several formats like mz5 [59], Toffee [60] and mzMLb [61] which are based on HDF5, which itself is a container format and can be considered the successor to netCDF.

Data standards in X-ray crystallography

Crystal structure analyses by X-ray diffraction are fundamental techniques in chemistry to determine the atomic and molecular structure of materials. These techniques measure the angles and intensities of a diffracted X-ray beam and calculate structural information from the data. In case of single-crystal measurements, the raw datasets can be very large, while other methods like powder X-ray diffraction produce only two-dimensional raw data. Therefore, data from the latter are exchanged in simple text files, exported from the instrument vendor software.

With the **Crystallographic Information File (CIF)**, there is a common exchange format for crystallographic data, which is developed and maintained by the International Union for Crystallography [43, 44]. The CIF is an implementation of the STAR file format and thus a text file which is organized in data blocks which are described by data names or tags. These data names are defined in plain text dictionaries, which use a controlled language and are readable for humans and computers. Besides the core dictionary with tags relevant for small-molecule and inorganic crystals, there are dictionaries for special applications like powder X-ray diffraction or macromolecular crystals (mmCIF). The possibility to extend the format by adding new dictionaries makes it ready for new methods and applications. Furthermore, with the Crystallographic Information Framework there is also a data model, which relies on the same principles as the file format and can be adapted to specific applications.

Databases and repositories like The Cambridge Structural Database (CSD) [62, 63] and Crystallography Open Database (COD) [64] will only accept CIF as format to deposit crystallographic data.

Data standards in X-ray absorption and fluorescence spectroscopy

X-ray absorption (XAS) as well as X-ray fluorescence (XRF) generates simple spectra described by the monochromatic X-ray radiation on the abscissa and the absorption of the sample on the ordinate. The spectra can be exported to in formats based on CSV, with multiple columns, but the units (*e.g.* energy, wavelength), column format and the included metadata often depend on the software used for measurement on the beamline or instrument, which interferes with interoperability. To compare XAS and XRF data measured on different beamlines and devices, it is also important to include parameters of the instrument and calibration into the dataset. For larger sets of XAS data the HDF5 format is considered as standard format [65], which is already used by some beamline software, such as BLISS on the ESRF [66].

For the interchange of single X-ray absorption spectra, the **XAFS Data Interchange (XDI)** format was proposed [67], which combines a dictionary of relevant metadata and the data table in a text file. Thus, it is readable for humans and computers and compatible with most of the existing software accepting x,y-tables. The authors of the format also provide an implementation in C and bindings for several other programming languages such as Fortran, Perl and Python. The format was already accepted for import to the reference sample database at Diamond Light Source [68] and the X-ray Absorption Data Library of the International X-ray Absorption Society [69].

There is no existing standard for XRF files, so most of the software tools provide some kind of import dialog to select matching columns and units from text files exported by the vendor software or can read data contained in HDF5 files.

Conclusion

Data standards are key to attain interoperable and reusable data as these standards do not solely facilitate data analysis and long-term archival but also interactions between scientists by publications and direct exchange. Versatile, robust and widely adopted data standards require an ecosystem of well-defined specifications, data models, examples and implementations. Efforts on standardization were and are driven by numerous organizations, consortia and communities. For a few methods specific data standards exist. For other methods, standards still need to be developed or existing standards need to be extended to also include essential method specific metadata.

Research funding: All authors, except Stefan Kuhn, acknowledge DFG (DOI: 10.13039/501100001659) funding under project number 441958208 (NFDI4Chem), Nicole Jung additionally acknowledges funding from Baden-Wuerttemberg Ministry of Science, Research and Art (DOI: 10.13039/501100003542) for the project Science Data Center MoMaF and Tobias Schulze acknowledges EU funding under grand agreement 733032 (HBM4EU).

References

[1] K. Rajan, H. O. Brinkhaus, A. Zielesny, C. Steinbeck. *J. Cheminf.* **12**, 60 (2020).

[2] M. D. Wilkinson, M. Dumontier, I. J. Jan Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons. *Sci. Data* **3**, 160018 (2016).

[3] T. Habermann. *Patterns (N Y)* **1**, 100004 (2020).

[4] M. J. Harvey, N. J. Mason, A. McLean, P. Murray-Rust, H. S. Rzepa, J. J. P. Stewart. *J. Cheminf.* **7**, 43 (2015).

[5] J. B. McAlpine, S.-N. Chen, A. Kutateladze, J. B. MacMillan, G. Appendino, B. Andersson, M. A. Beniddir, M. W. Biavatti, S. Bluml, A. Boufridi, M. S. Butler, R. J. Capon, Y. H. Choi, D. Coppage, P. Crews, M. T. Crimmins, M. Csete, P. Dewapriya, J. M. Egan, M. J. Garson, G. Genta-Jouve, W. H. Gerwick, H. Gross, M. K. Harper, P. Hermanto, J. M. Hook, L. Hunter, D. Jeannerat, N.-Y. Ji, T. A. Johnson, D. G. I. Kingston, H. Koshino, H.-W. Lee, L. Guy, J. Li, R. G. Linington, M. Liu, K. L. McPhail, T. F. Molinski, B. S. Moore, J.-W. Nam, R. P. Neupane, M. Niemitz, J.-M. Nuzillard, N. H. Oberlies, F. M. M. Ocampos, G. Pan, R. J. Quinn, D. S. Reddy, J.-H. Renault, J. Rivera-Chávez, W. Robien, C. M. Saunders, T. J. Schmidt, C. Seger, B. Shen, C. Steinbeck, H. Stuppner, S. Sturm, O. Taglialatela-Scafati, D. J. Tantillo, R. Verpoorte, B.-G. Wang, C. M. Williams, P. G. Williams, J. Wist, J.-M. Yue, C. Zhang, Z. Xu, C. Simmler, D. C. Lankin, J. Bisson, G. F. Pauli. *Nat. Prod. Rep.* **36**, 35 (2019).

[6] L. M. J. Kroon-Batenburg, J. R. Helliwell, B. McMahon, T. C. Terwilliger. *IUCrJ* **4**, 87 (2017).

[7] T. Miyakawa. *Mol. Brain* **13**, 24 (2020).

[8] *rfc4180*. WEBSITE. URL: https://datatracker.ietf.org/doc/html/rfc4180 (visited Oct 29, 2021).

[9] *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. WEBSITE. URL: https://www.w3.org/%7BTR%7D/xml/ (visited Oct 29, 2021).

[10] *ECMA-404 – Ecma International*. WEBSITE. URL: https://www.ecma-international.org/publications-and-standards/standards/ecma-404/ (visited Oct 29, 2021).

[11] S. Soiland-Reyes, P. Sefton, M. Crosas, L. Jael Castro, F. Coppens, J. M. Fernéndez, D. Garijo, B. Grüning, M. La Rosa, S. Leo, E. Ó. Carrag'ain, M. Portier, A. Trisovic, RO-Crate Community, P. Groth, C. Goble. *Data Sci.* **Preprint**, 1 (2022).

[12] *International Union of Pure and Applied Chemistry*. WEBSITE. URL: https://iupac.org/ (visited Jul 05, 2021).

[13] R. Cammack, F. Yang, R. J. Lancashire, J. P. Maher, P. S. McIntyre, R. Morse. *Pure Appl. Chem.* **78**, 613 (2006).

[14] M. Todd, A. J. Jarrett, N. Young, D. E. Vanderwall, D. Della Corte. *Drug Discov. Today* **26**, 1922 (2021).

[15] *Unified Data Model – Pistoia Alliance*. WEBSITE. URL: https://www.pistoiaalliance.org/projects/current-projects/unified-data-model/ (visited Oct 29, 2021).

[16] H. Bär, R. Hochstrasser, B. Papenfub. *J. Lab. Autom.* **17**, 86 (2012).

[17] C. Steinbeck, O. Koepler, F. Bach, S. Herres-Pawlis, N. Jung, J. Liermann, S. Neumann, M. Razum, C. Baldauf, B. Frank, T. Bocklitz, F. Boehm, B. Frank, C. Paul, T. Engel, M. Hicks, S. Kast, C. Kettner, W. Koch, G. Lanza, A. Link, R. Mata, W. Nagel, A. Porzel, N. Schlörer, T. Schulze, H.-G. Weinig, W. Wenzel, L. Wessjohann, S. Wulle. *RIO* **6**, e55852 (2020).

[18] *GO FAIR initiative: Make your data & services FAIR*. WEBSITE. URL: https://www.go-fair.org/ (visited Jul 05, 2021).

[19] *Home – CODATA, The Committee on Data for Science and Technology*. WEBSITE. URL: https://codata.org/ (visited Jul 05, 2021).

[20] F. Berman, M. Crosas. *Harv. Data Sci. Rev.* **2** (2020), https://doi.org/10.1162/99608f92.5e126552.

[21] *FORCE11 |The future of research communications and e-scholarship*. WEBSITE. URL: https://www.force11.org/ (visited Jul 05, 2021).

[22] R. E. Duerr, R. R. Downs, C. Tilmes, B. Barkstrom, W. Christopher Lenhardt, G. Joseph, L. E. Bermudez, P. Slaughter. *Earth Sci. Inf.* **4**, 139 (2011).

[23] J. Neumann, B. Jan. *J. Comput. Aided Mol. Des.* **28**, 1035 (2014).

[24] B. Jan. DataCite – a global registration agency for research data. In 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, pp. 257–261, IEEE, Beijing, China (2009). URL: http://ieeexplore.ieee.org/document/5361881/ (visited Oct 06, 2021).

[25] S. Susanna-Assunta, P. McQuilton, P. Rocca-Serra, A. Gonzalez-Beltran, M. Izzo, A. L. Lister, M. Thurston, FAIRsharing Community. *Nat. Biotechnol.* **37**, 358 (2019).

[26] H. Pampel, V. Paul, S. Frank, B. Roland, M. Kindling, J. Klump, H.-J. Goebelbecker, J. Gundlach, P. Schirmbacher, U. Dierolf. *PLoS One* **8**, e78080 (2013).

[27] R. S. McDonald, P. A. Wilks. *Appl. Spectrosc.* **42**, 151 (1988).

[28] A. N. Davies, P. Lampen. *Appl. Spectrosc.* **47**, 1093 (1993).

[29] P. Lampen, H. Heinrich, A. N. Davies, M. Linscheid. *Appl. Spectrosc.* **48**, 1545 (1994).

[30] *OpenLab/AnIML_python GitLab*. WEBSITE. URL: https://gitlab.com/opensourcelab/animl%5C_python (visited Oct 29, 2021).

[31] R. Rew, G. Davis. *IEEE Comput. Graph. Appl.* **10**, 76 (1990).

[32] S. Susanna-Assunta, P. Rocca-Serra, D. Field, E. Maguire, C. Taylor, O. Hofmann, H. Fang, S. Neumann, W. Tong, L. Amaral-Zettler, K. Begley, T. Booth, L. Bougueleret, G. Burns, B. Chapman, T. Clark, L.-A. Coleman, J. Copeland, S. Das, A. de Daruvar, P. de Matos, I. Dix, E. Scott, C. T. Evelo, M. J. Forster, P. Gaudet, J. Gilbert, C. Goble, J. L. Griffin, D. Jacob, J. Kleinjans, H. Lee, K. Haug, H. Hermjakob, S. J. H. Sui, A. Laederach, S. Liang, S. Marshall, A. McGrath, E. Merrill, D. Reilly, M. Roux, C. E. Shamu, C. A. Shang, C. Steinbeck, A. Trefethen, B. Williams-Jones, K. Wolstencroft, I. Xenarios, W. Hide. *Nat. Genet.* **44**, 121 (2012).

[33] S. Susanna-Assunta, P. Rocca-Serra, A. Gonzalez-Beltran, D. Johnson, ISA Community. *Zenodo* (2016), https://doi.org/10.5281/zenodo.163640.

[34] A. Roth, R. Jopp, R. Schäfer, G. W. Kramer. *JALA: J. Assoc. Lab. Autom.* **11**, 247 (2006).

[35] B. A. Schäfer, D. Poetz, G. W. Kramer. *JALA: J. Assoc. Lab. Autom.* **9**, 375 (2004).

[36] D. Johnson, D. Batista, K. Cochrane, R. P. Davey, E. Anthony, A. Gonzalez-Beltran, K. Haug, M. Izzo, M. Larralde, T. N. Lawson, A. Minotto, P. Moreno, V. Chandrasekhar Nainala, C. O'Donovan, L. Pireddu, P. Roger, F. Shaw, C. Steinbeck, R. J. M. Weber, S. Susanna-Assunta, P. Rocca-Serra. *Gigascience* **10**, giab060 (2021).

[37] M. Turewicz, E. W. Deutsch. *Methods Mol. Biol.* **696**, 179 (2011).

[38] E. W. Deutsch. *Methods Mol. Biol.* **604**, 319 (2010).

[39] B. Erickson. *Anal. Chem.* **72**, 103 A (2000).

[40] M. Larralde, T. N. Lawson, R. J. M. Weber, P. Moreno, K. Haug, P. Rocca-Serra, M. R. Viant, C. Steinbeck, R. M. Salek. *Bioinformatics* **33**, 2598 (2017).

[41] M. Pupier, J.-M. Nuzillard, J. Wist, N. E. Schlörer, S. Kuhn, M. Erdelyi, C. Steinbeck, A. J. Williams, C. Butts, T. D. W. Claridge, B. Mikhova, W. Robien, H. Dashti, H. R. Eghbalnia, C. Farès, C. Adam, P. Kessler, F. Moriaud, M. Elyashberg, D. Argyropoulos, M. Pérez, P. Giraudeau, R. R. Gil, T. Paul, D. Jeannerat. *Magn. Reson. Chem.* **56**, 703 (2018).

[42] S. Kuhn, L. H. E. Wieske, T. Paul, D. Schober, N. E. Schlörer, J.-M. Nuzillard, P. Kessler, J. Junker, A. Herráez, C. Farès, M. Erdélyi, D. Jeannerat. *Magn. Reson. Chem.* **59**, 792 (2021).

[43] I. D. Brown. *J. Res. Natl. Inst. Stand. Technol.* **101**, 341 (1996).

[44] I. D. Brown, B. McMahon. *Acta Crystallogr. B Struct. Sci. Crystallogr. Eng. Mater.* **58**, 317 (2002).

[45] M. Alexander Ruhl, R. Schäfer, G. W. Kramer. *SpectroML: An Extensible Markup Language for the Interchange of Molecular Spectrometry Data*. Tech. rep., NIST (National Institute of Standards and Technology), Gaithersburg, MD (2002). URL: https://nvlpubs.nist.gov/nistpubs/Legacy/%7BIR%7D/nistir6821.pdf (visited Jul 19, 2021).

[46] A. D. Thi Nguyen, A. Arslan, J. Travis, M. Smith, R. Schafer, G. W. Kramer. *JALA: J. Assoc. Lab. Autom.* **9**, 346 (2004).

[47] J. Bisson, C. Simmler, S.-N. Chen, J. B. Friesen, D. C. Lankin, J. B. McAlpine, G. F. Pauli. *Nat. Prod. Rep.* **33**, 1028 (2016).

[48] P. Lampen, J. Lambert, R. J. Lancashire, R. S. McDonald, P. S. McIntyre, D. N. Rutledge, T. Fröhlich, A. N. Davies. *Pure Appl. Chem.* **71**, 1549 (1999).

[49] D. Schober, D. Jacob, M. Wilson, J. A. Cruz, A. Marcu, J. R. Grant, A. Moing, C. Deborde, L. F. de Figueiredo, K. Haug, P. Rocca-Serra, J. Easton, T. M. D. Ebbels, J. Hao, C. Ludwig, U. L. Günther, A. Rosato, M. S. Klein, I. A. Lewis, C. Luchinat, A. R. Jones, A. Grauslys, M. Larralde, M. Yokochi, N. Kobayashi, A. Porzel, J. L. Griffin, M. R. Viant, D. S. Wishart, C. Steinbeck, R. M. Salek, S. Neumann. *Anal. Chem.* **90**, 649 (2018).

[50] N. S. Kale, K. Haug, P. Conesa, K. Jayseelan, P. Moreno, P. Rocca-Serra, V. Chandrasekhar Nainala, R. A. Spicer, M. Williams, X. Li, R. M. Salek, J. L. Griffin, C. Steinbeck. *Curr. Protoc. Bioinf.* **53**, 14.13.1 (2016).

[51] A. Gutmanas, P. D. Adams, B. Bardiaux, H. M. Berman, D. A. Case, R. H. Fogh, P. Güntert, P. M. S. Hendrickx, T. Herrmann, G. J. Kleywegt, N. Kobayashi, O. F. Lange, J. L. Markley, G. T. Montelione, M. Nilges, T. J. Ragan, C. D. Schwieters, R. Tejero, E. L. Ulrich, S. Velankar, W. F. Vranken, J. R. Wedell, J. Westbrook, D. S. Wishart, G. W. Vuister. *Nat. Struct. Mol. Biol.* **22**, 433 (2015).

[52] E. L. Ulrich, K. Baskaran, H. Dashti, Y. E. Ioannidis, M. Livny, P. R. Romero, D. Maziuk, J. R. Wedell, H. Yao, H. R. Eghbalnia, J. C. Hoch, J. L. Markley. *J. Biomol. NMR* **73**, 5 (2019).

[53] W. F. Vranken, W. Boucher, T. J. Stevens, R. H. Fogh, A. Pajon, M. Llinas, E. L. Ulrich, J. L. Markley, J. Ionides, E. D. Laue. *Proteins* **59**, 687 (2005).

[54] *Mascot database search |Data file format for mass spectrometry peak lists*. WEBSITE. URL: https://newbsrcmascot.st-andrews.ac.uk/mascot/help/data%5C_file%5C_help.html (visited Oct 28, 2021).

[55] Users Guide. *NIST Mass Spectral Search Program (Version 2.0g)*, NIST (National Institute of Standards and Technology), Gaithersburg, MD (2011).

[56] S. M. Lin, L. Zhu, A. Q. Winter, M. Sasinowski, W. A. Kibbe. *Expert Rev. Proteomics* **2**, 839 (2005).

[57] S. Orchard, L. Montechi-Palazzi, E. W. Deutsch, P.-A. Binz, A. R. Jones, N. Paton, A. Pizarro, D. M. Creasy, J. Wojcik, H. Hermjakob. *Proteomics* **7**, 3436 (2007).

[58] J. Griss, F. Reisinger, H. Hermjakob, J. Antonio Vizcaíno. *Proteomics* **12**, 795 (2012).

[59] M. Wilhelm, M. Kirchner, J. A. J. Steen, H. Steen. *Mol. Cell. Proteomics* **11**, O111.011379 (2012).

[60] B. Tully. *Sci. Rep.* **10**, 8939 (2020).

[61] R. S. Bhamber, A. Jankevics, E. W. Deutsch, A. R. Jones, A. W. Dowsey. *J. Proteome Res.* **20**, 172 (2021).

[62] C. R. Groom, I. J. Bruno, M. P. Lightfoot, S. C. Ward. *Acta Crystallogr. B Struct. Sci. Crystallogr. Eng. Mater.* **72**, 171 (2016).

[63] R. Taylor, P. A. Wood. *Chem. Rev.* **119**, 9427 (2019).

[64] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, M. Peter, R. T. Downs, A. Le Bail. *Nucleic Acids Res.* **40**, D420 (2012).

[65] B. Ravel, J. R. Hester, V. A. Solé, M. Newville. *J. Synchrotron Radiat.* **19**, 869 (2012).

[66] M. Guijarro, A. Beteva, T. Coutinho, M.-C. Dominguez, C. Guilloud, A. Homs, J. Meyer, V. Michel, E. Papillon, M. Perez, S. Ã. Petitdemange. BLISS – experiments control for ESRF EBS beamlines, pp. 1060–1066, JACoW Publishing, Geneva, Switzerland (2018).

[67] B. Ravel, M. Newville. *J. Phys. Conf. Ser.* **712**, 012148 (2016).

[68] G. Cibin, D. Gianolio, S. A. Parry, S. Tom, O. Moore, R. Draper, L. A. Miller, T. Alexander, C. L. Doswell, A. Graham. *Radiat. Phys. Chem.* **175**, 108479 (2020).

[69] *XASLIB: X-ray Absorption Data Library*. WEBSITE. URL: https://xaslib.xrayabsorption.org/elem/ (visited Oct 18, 2021).