A universal standard archive file for adsorption data

Jack D. Evans,* Volodymyr Bon, Irena Senkovska, and Stefan Kaskel

Department of inorganic chemistry Technische Universität Dresden Bergstraße 66, 01062 Dresden, Germany

E-mail: jack.evans@tu-dresden.de

Abstract

The rise in interest of new advanced adsorbents for application in promising areas has followed a lack of detailed and open reporting of experimental or even computational studies. This report details the specification of a new standard adsorption information file (AIF) inspired by the ubiquitous crystallographic information file (CIF) and based on the self-defining text archive and retrieval (STAR) procedure, also used to represent biological nuclear magnetic resonance experiments (NMR-STAR). The AIF is a flexible, general and easily extended free-format archive file and is readily human and machine readable— simple to edit using a basic text editor or parse for database curation. This format represents the first steps to an open adsorption data format. An open format sorely needed for the electronic transmission of adsorption data between laboratories, journals and larger databases in the effort to increase open science in the field of porous materials.

Introduction

It is important in many branches of science for there to exist a uniform and flexible method of archiving and exchanging data. The data presented in research contributions must be readily accessible and easily verifiable. Key to scientific discovery and innovation the use of good data management practices can provide significant knowledge through the reuse of original data by the community after the data publication process.[1] Unfortunately, the existing ecosystem surrounding scholarly data publication often prevents the research community extracting maximum benefit from published reports. This led to the foundational principles as discussed by Wilkinson et al.— Findability, Accessibility, Interoperability, and Reusability— that serve as a guideline to to maximize the added-value gained by digital publishing.[2]

Porous solids are of important scientific and industrial interest owing to their impressive ability to interact with all manner of atoms, ions and molecules throughout the entirety of the solid, not only the surface.[3] This has led to many exciting applications for porous materials, including ion exchange, gas storage, separations and catalysis.[4] Measuring the interaction of gas with a material is crucial to understanding and comparing different porous materials. The quantity of gas adsorbed is measured in any convenient units, but for the presentation of the data, the International Union of Pure and Applied Chemistry (IUPAC) has recommended that the amount adsorbed should be expressed in moles per gram of outgassed adsorbent.[5] To facilitate the comparison of adsorption data, the IUPAC also recommended that adsorption isotherms are displayed in a graphical form with the amount adsorbed (preferably presented using $\mathrm{mol\,g^{-}1}$ plotted against the equilibrium relative pressure (p/p_0), where p_0 is the saturation pressure of the pure adsorptive at the experiment temperature, or when the temperature is above the critical temperature of the adsorbing gas, against absolute pressure.

Rapid advances in computer simulations, coupled with the expansion of fruitful local, national and international networks, have fueled the need for a standardization of adsorption data beyond the current IUPAC recommendations. The variety and relative inflexibility of existing adsorption data formats and conventional graphical reporting inhibits the effective use of reported data. A general, flexible, rapidly extensible and universal file format protocol following the guidelines presented by Wilkinson et al. is essential to further accelerate the development of porous materials. As with other standard formats, this must be machine-independent and portable so that accessibility to data items is independent of their point of origin. It must also allow new data items to be incorporated without the need to modify existing files to enable an evolving format.

In this submission we outline our proposal for a new standard adsorption information file (AIF) and the clear advantages of this format compared to the currently accepted graphical deposition method. The AIF is free-format archive file that is readily human and machine readable, both simple to edit using a basic text editor or parse using computer programs. This represents the first flexible, general and easily extended format for representing the results of adsorption experiments and the first steps towards standardizing a format for archiving adsorption data.

AIF structure and syntax

The information contained within an AIF is arranged in a strict structural arrangement. Alternative data storage structures were considered, during the development the AIF, such as the comma-separated-value (CSV) file. The key, however, is that a universal data format must also include a well-structured record of metadata associated with the adsorption experiment. Without the metadata of, containing at the very least, what gas or temperature was used during the experiment the resulting adsorption data is meaningless. Metadata is also important in the development of databases. This is crucial to identify identical or related data, understand the evolution of databases over time, and also permit further examination for specific data items. As a result the STAR data structure was adopted,[6] which is famously used in the crystallographic information file (CIF) and biological nuclear magnetic resonance experiments (NMR-STAR).[7,8] An important property of the STAR format is that its syntax is defined by a few simple rules using an object/relational model. The data stored in a STAR file is self-descriptive as the underpinning data structure consists of tag-value pairs or data item. This data item is the combination a data value with the data name, an identifying unique tag. This approach is flexible and able to represent. nested and repetitive data, which are particularly important for the collection of pressure and adsorbed amount that comprise an adsorption isotherm experiment.

The data structure of an AIF is comprised of comment lines and data blocks, which contains the required metadata and quantities related to the adsorption experiment, separated for adsorption and desorption (Figure 1). Files can contain any number of data blocks to represent multiple experiments. Fundamentally, a data block is a data cell containing a sequence of data items, data loops. It starts with a data block code statement (`data_example`) where a block code is a unique identifying code with in a file. A data block is closed by another either another data block code statement or the end of file. Comment lines can be placed anywhere in the file and are identified by a hash (`#`) symbol. As these comment lines do not necessarily represent structured data, which is able to be easily parsed, it is not recommended to solely include important experimental conditions in this way. Single values or strings can be included within a data block using the data name and data value pairs. Data values are preceded by an identifying data name that is identify by an underscore character, for example `_temperature`. A data loop structure consists of a loop statement (`loop_`) followed by a list of data names and then a list of data, each of which contains data values that matches the data names in the list of data names. Though data loops can be nested at any level this is not necessary for the current implementation. Importantly, data values that follow the nested data declarations must be in exact multiples of the number of data names (which are referred to as loop packets). The data loops, as used here, are terminated by a new data item, a new data loop or an end of file. These data loops can thus straightforwardly represent the data recorded at each point of an adsorption and desorption experiment, for example the pressure and quantities adsorbed.

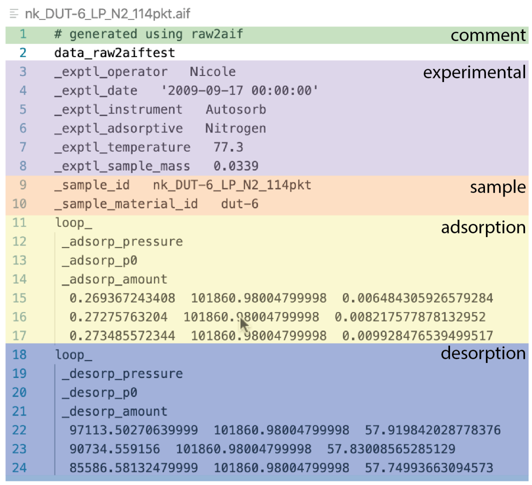

Figure 1: The structure of a typical adsorption information file (AIF) with the different data areas representing metadata (experimental and sample information) and the adsorption and desorption data.

The above data structures of the STAR format prove ideal for describing the detailed quantities relating to an adsorption experiment. An entire experiment can be described within a single data block. Initially, data name and data value pairs can be used to identify quantities relating to the whole experiment such as the date and time, temperature and adsorptive. The data points for the adsorption and desorption branch can then be detailed in separate data loops. For each point, quantities such as the absolute pressure, relative pressure and amount adsorbed can be detailed. This leads us to the general data structure for the AIF (Figure 1).

AIF dictionary

Each data item within a data block is identified using a unique data name. The currently accepted AIF data names are listed and defined in Table 1. These data items form a first version of a core dictionary, which are intended for primarily reporting primary physisorption data for gas on porous materials. As a result the dictionary contains information about crucial experimental parameters, basic sample characterization, adsorption and desorption measures, each distinguished by a data name prefix (`_exptl_`, `_sample_`, `_adsorp_` and `_desorp_`).

Table 1: Core dictionary of data items

data name	description
`_exptl_operator`	name of the person who ran the experiment (string)
`_exptl_date`	date and time of the experiment (date-time string representation)
`_exptl_instrument`	instrument id used for the experiment (string)
`_exptl_adsorptive`	name of the adsorptive (string)
`_exptl_temperature`	temperature of the experiment (float)
`_exptl_sample_mass`	mass of the sample (float, gram)
`_sample_id`	unique identifying code used by the operator (string)
`_sample_material_name`	designated name for the material (string)
`_adsorp_pressure`	pressure of the adsorption measurement (float, pascal)
`_adsorp_p0`	saturation pressure of the adsorption measurement (float, pascal)
`_adsorp_amount`	amount adsorbed during the adsorption measurement (float, $\mathrm{mol\,kg^{-1}}$)
`_desorp_pressure`	pressure of the desorption measurement (float, pascal)
`_desorp_p0`	saturation pressure of the desorption measurement (float, pascal)
`_desorp_amount`	amount adsorbed during the desorption measurement (float, $\mathrm{mol\,kg^{-1}}$)

The goal of the experimental data names is to capture the information required to accurately portray the results of experiment, which at the very minimum are the temperature and adsorptive. As metadata should include information that may aid in identifying possible errors this also includes the mass of sample used in the experiment. Additionally, data is included relative to the date, time and person responsible for the measurement to produce a general picture of the provenance of the experiment.

Details of the sample are recorded for comparison and to uniquely identify the sample that was used in the experiment. Simply, this includes the sample identity code (`_sample_id`), a laboratory code used to record and represent the sample and a string relating to the general designation given to the material (`_sample_material_name`).

For each data point measured during an adsorption experiments data can be stored that relates to the equilibrated pressure, the saturation pressure and the amount adsorbed.

Importantly, these should be stored in units of pascal and $\mathrm{mol\,kg^{-1}}$ to follow the IUPAC recommendations.[5] The saturation pressure can be recorded at each data point as we have discovered that some instruments will calculate and report this value for every point.

The above features serve as an example of the data that can be archived in a AIF and can be extended in future versions. For example, the intention of this initial version is to provide digital archive of the presented primary data of an adsorption experiment, the isotherm. As a result, it describes only basic information that usually exported from adsorption instruments and a few convenient metadata items. The flexibility of this approach can be extended in future versions to include more detailed parameters not commonly presented, such as dead volume measurements, or combined with CIF data names to archive in situ measurements.

Creating and using the AIF

A simple python 3.7 routine was produced during the development of the AIF to convert the plain text data output from Quantachrome and Belsorb instruments, used in our lab and many others. Specifically, this has been written to parse raw analysis text files exported by Quantachrome software and the raw data files (*.DAT*) and CSV files exported by Belsorb software. At this moment we are unable to read binary data files that are saved by Quatachrome instruments (*.qps*).

This simple program will parse the output data to find data items detailed in the AIF dictionary. These data items are converted to the appropriate units and the isotherm is split into the adsorption and desorption parts. Subsequently an AIF is produced using the pymatgen CIF routines.[9] This demonstrated that routines and methods developed for treating CIF files can be used in this application to adsorption data. Furthermore, produced AIF files can also be read using existing CIF methods, such as the `gemmi` python library that is considered the the fastest open-source CIF parser.[10] We have used `gemmi` library to demonstrate how an isotherm can be displayed from an AIF using a few lines. This is a very simple implementation and it is hoped that future versions, following the cooperation of several manufactures, could result in to routines to parse the information directly from binary data used by the instruments.

To illustrate the importance of a universal data format for adsorption data we can compare data contained within an AIF to data digitized from a figure in a manuscript, for the same experiment. The porous material DUT-6 (also referred to as MOF-206) is a metal-organic framework produced from zinc nitrate and two different different ligands; 2,6-naphthalenedicarboxylic acid and benzene-1,3,5-tribenzoic acid.[11] DUT-6 has mesopores with dodecahedral geometry and smaller cages that produce an extremely porous material with large adsorption capacities for$n$-butane (up to 1.1 $\mathrm{g\,g^{-1}}$ at 293 K), methane (230 $\mathrm{mg\,g^{-1}}$ at 100 bar and 298 K) and hydrogen (60 $\mathrm{mg\,g^{-1}}$ at 50 bar and 77 K). The nitrogen adsorption isotherm at 77 K was reported in the original manuscript using relative pressure and a linear axis (Figure /reffgr:example). From this plot the isotherm can be classified as type Ib, clearly exhibiting characteristics of a hierarchal pore system with micropores and mesopores. There is evidently a distinct step $p/p_0 = 0.17$, however, from this plot it is not possible to correctly identify low pressure features. The low pressure regions is important for understanding the process of initial pore filling or for independent application of analysis methods, such as the Brunauer–Emmett–Teller (BET) method.[12] The reported isotherm was digitized using WebPlotDigitizer, which uses a time-consuming but powerful point-picking methodology.[13] In this methodology, each point in the graph is identified by clicking on the image area and the point's location in on-screen pixels is then converted to data values, based on an axes calibration. The digitized isotherm and the data from the experiment, when plotted on semi-log axes, clearly shows important low pressure information, such as the point of initial pore-filling, is lost when this information is simply included as a figure in a manuscript (Figure 2).

To prevent this information loss and to allow further analysis of reported adsorption experiments by other researchers, adsorption data should be archived and shared using a universal and complete data format like that of the AIF.

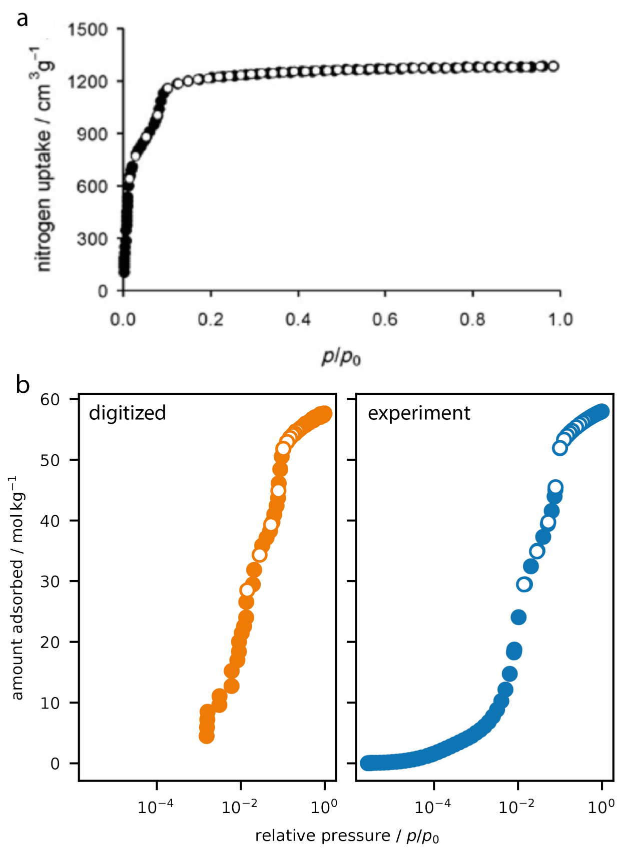

Figure 2: The reported nitrogen adsorption isotherm of DUT-6 recorded at 77 K (a) as reported in Ref. 11. A digitized representation of this reported isotherm was achieved using a point-picking method (b) and compared to the experimental data as contained within an AIF (c).

Summary and perspective

The rise in interest of new advanced adsorbents for application in promising areas has followed a lack of detailed and open reporting of experimental or even computational studies. This report details the specification of a new standard adsorption information file (AIF) inspired by the ubiquitous crystallographic information file (CIF) and based on the self-defining text archive and retrieval (STAR) procedure, also used to represent biological nuclear magnetic resonance experiments (NMR-STAR). The AIF is a flexible, general and easily extended free-format archive file and is readily human and machine readable— simple to edit using a basic text editor or parse for database curation. This format represents the first steps to an open adsorption data format. An open format sorely needed for the electronic transmission of adsorption data between laboratories, journals and larger databases in the effort to increase open science in the field of porous materials.

Acknowledgement

J. D. E. acknowledges the support of the Alexander von Humboldt foundation and HPC platforms provided by a GENCI grant (A0070807069) and the Center for Information Services and High Performance Computing (ZIH) at TU Dresden.

References

(1) Coudert, F.-X. Materials Databases: The Need for Open, Interoperable Databases with Standardized Data and Rich Metadata. *Advanced Theory and Simulations* **2019**, *2*, 1900131.

(2) Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **2016**, *3*.

(3) Davis, M. E. Ordered porous materials for emerging applications. *Nature* **2002**, *417*, 813–821.

(4) Sun, M.-H.; Huang, S.-Z.; Chen, L.-H.; Li, Y.; Yang, X.-Y.; Yuan, Z.-Y.; Su, B.-L. Applications of hierarchically structured porous materials from energy storage and conversion, catalysis, photocatalysis, adsorption, separation, and sensing to biomedicine. *Chemical Society Reviews* **2016**, *45*, 3479–3563.

(5) Thommes, M.; Kaneko, K.; Neimark, A. V.; Olivier, J. P.; Rodriguez-Reinoso, F.;

Rouquerol, J.; Sing, K. S. Physisorption of gases, with special reference to the evaluation of surface area and pore size distribution (IUPAC Technical Report). *Pure and Applied Chemistry* **2015**, *87*, 1051–1069.

(6) Hall, S. R.; Spadaccini, N. The STAR File: detailed specifications. *Journal of Chemical Information and Modeling* **1994**, *34*, 505–508.

(7) Hall, S. R.; Allen, F. H.; Brown, I. D. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A Foundations of Crystallography* **1991**, *47*, 655–685.

(8) Ulrich, E. L.; Baskaran, K.; Dashti, H.; Ioannidis, Y. E.; Livny, M.; Romero, P. R.; Maziuk, D.; Wedell, J. R.; Yao, H.; Eghbalnia, H. R.; Hoch, J. C.; Markley, J. L. NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments. *Journal of Biomolecular NMR* **2018**, *73*, 5–9.

(9) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **2013**, *68*, 314–319.

(10) Global Phasing Ltd,; CCP4, GEMMI - library for structural biology. 2020; `https://github.com/project-gemmi/gemmi`.

(11) Klein, N.; Senkovska, I.; Gedrich, K.; Stoeck, U.; Henschel, A.; Mueller, U.; Kaskel, S. A mesoporous metal-organic framework. *Angewandte Chemie International Edition* **2009**, *48*, 9954–9957.

(12) Rouquerol, J.; Llewellyn, P.; Rouquerol, F. *Studies in Surface Science and Catalysis*; Elsevier, 2007; pp 49–56.

(13) Rohatgi, A. WebPlotDigitizer. 2020; https://automeris.io/WebPlotDigitizer.

Graphical TOC Entry

The surrounding frame is 9 cm by 3.5 cm, which is the maximum permitted for *Journal of the American Chemical Society* graphical table of content entries. The box will not resize if the content is too big: instead it will overflow the edge of the box.