A universal standard archive file for adsorption data

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

1

*Department of inorganic chemistry Technische Universit¨at Dresden Bergstraße 66, 01062 Dresden, Germany*

E-mail: [jack.evans@tu-dresden.de](mailto:jack.evans@tu-dresden.de)

2 **Abstract**

3 The rise in interest of new advanced adsorbents for application in promising areas

4 has followed a lack of detailed and open reporting of experimental or even computa-

5 tional studies. This report details the specification of a new standard adsorption infor-

6 mation file (AIF) inspired by the ubiquitous crystallographic information file (CIF) and

7 based on the self-defining text archive and retrieval (STAR) procedure, also used to

8 represent biological nuclear magnetic resonance experiments (NMR-STAR). The AIF

9 is a flexible, general and easily extended free-format archive file and is readily human

10 and machine readable— simple to edit using a basic text editor or parse for database

11 curation. This format represents the first steps to an open adsorption data format. An

12 open format sorely needed for the electronic transmission of adsorption data between

13 laboratories, journals and larger databases in the effort to increase open science in the

14 field of porous materials.

# 15 Introduction

16 It is important in many branches of science for there to exist a uniform and flexible method of

17 archiving and exchanging data. The data presented in research contributions must be readily

18 accessible and easily verifiable. Key to scientific discovery and innovation the use of good data

19 management practices can provide significant knowledge through the reuse of original data by

20 the community after the data publication process.1 Unfortunately, the existing ecosystem

21 surrounding scholarly data publication often prevents the research community extracting

22 maximum benefit from published reports. This led to the foundational principles as discussed

23 by Wilkinson et al.— Findability, Accessibility, Interoperability, and Reusability— that serve

24 as a guideline to to maximize the added-value gained by digital publishing.2

25 Porous solids are of important scientific and industrial interest owing to their impressive

26 ability to interact with all manner of atoms, ions and molecules throughout the entirety

27 of the solid, not only the surface.3 This has led to many exciting applications for porous

28 materials, including ion exchange, gas storage, separations and catalysis.4 Measuring the

29 interaction of gas with a material is crucial to understanding and comparing different porous

30 materials. The quantity of gas adsorbed is measured in any convenient units, but for the

31 presentation of the data, the International Union of Pure and Applied Chemistry (IUPAC)

32 has recommended that the amount adsorbed should be expressed in moles per gram of

33 outgassed adsorbent.5 To facilitate the comparison of adsorption data, the IUPAC also rec-

34 ommended that adsorption isotherms are displayed in a graphical form with the amount

35 adsorbed (preferably presented using mol g*−*1 plotted against the equilibrium relative pres-

36 sure (*p/p*0), where *p*0 is the saturation pressure of the pure adsorptive at the experiment

37 temperature, or when the temperature is above the critical temperature of the adsorbing

38 gas, against absolute pressure.

39 Rapid advances in computer simulations, coupled with the expansion of fruitful local,

40 national and international networks, have fueled the need for a standardization of adsorption

41 data beyond the current IUPAC recommendations. The variety and relative inflexibility of

42 existing adsorption data formats and conventional graphical reporting inhibits the effective

43 use of reported data. A general, flexible, rapidly extensible and universal file format protocol

44 following the guidelines presented by Wilkinson et al. is essential to further accelerate the

45 development of porous materials. As with other standard formats, this must be machine-

46 independent and portable so that accessibility to data items is independent of their point

47 of origin. It must also allow new data items to be incorporated without the need to modify

48 existing files to enable an evolving format.

49 In this submission we outline our proposal for a new standard adsorption information file

50 (AIF) and the clear advantages of this format compared to the currently accepted graphical

51 deposition method. The AIF is free-format archive file that is readily human and machine

52 readable, both simple to edit using a basic text editor or parse using computer programs.

53 This represents the first flexible, general and easily extended format for representing the

54 results of adsorption experiments and the first steps towards standardizing a format for

55 archiving adsorption data.

# 56 AIF structure and syntax

57 The information contained within an AIF is arranged in a strict structural arrangement.

58 Alternative data storage structures were considered, during the development the AIF, such

59 as the comma-separated-value (CSV) file. The key, however, is that a universal data format

60 must also include a well-structured record of metadata associated with the adsorption exper-

61 iment. Without the metadata of, containing at the very least, what gas or temperature was

62 used during the experiment the resulting adsorption data is meaningless. Metadata is also

63 important in the development of databases. This is crucial to identify identical or related

64 data, understand the evolution of databases over time, and also permit further examina-

65 tion for specific data items. As a result the STAR data structure was adopted,6 which is

66 famously used in the crystallographic information file (CIF) and biological nuclear magnetic

67 resonance experiments (NMR-STAR).7,8 An important property of the STAR format is that

68 its syntax is defined by a few simple rules using an object/relational model. The data stored

69 in a STAR file is self-descriptive as the underpinning data structure consists of tag-value

70 pairs or data item. This data item is the combination a data value with the data name, an

71 identifying unique tag. This approach is flexible and able to represent. nested and repetitive

72 data, which are particularly important for the collection of pressure and adsorbed amount

73 that comprise an adsorption isotherm experiment.

74 The data structure of an AIF is comprised of comment lines and data blocks, which con-

75 tains the required metadata and quantities related to the adsorption experiment, separated

76 for adsorption and desorption (Figure 1). Files can contain any number of data blocks to

77 represent multiple experiments. Fundamentally, a data block is a data cell containing a se-

78 quence of data items, data loops. It starts with a data block code statement (data example)

79 where a block code is a unique identifying code with in a file. A data block is closed by

80 another either another data block code statement or the end of file. Comment lines can

81 be placed anywhere in the file and are identified by a hash (#) symbol. As these comment

82 lines do not necessarily represent structured data, which is able to be easily parsed, it is not

83 recommended to solely include important experimental conditions in this way. Single values

84 or strings can be included within a data block using the data name and data value pairs.

85 Data values are preceded by an identifying data name that is identify by an underscore

86 character, for example temperature. A data loop structure consists of a loop statement

87 (loop ) followed by a list of data names and then a list of data, each of which contains data

88 values that matches the data names in the list of data names. Though data loops can be

89 nested at any level this is not necessary for the current implementation. Importantly, data

90 values that follow the nested data declarations must be in exact multiples of the number

91 of data names (which are referred to as loop packets). The data loops, as used here, are

92 terminated by a new data item, a new data loop or an end of file. These data loops can thus

93 straightforwardly represent the data recorded at each point of an adsorption and desorption

94 experiment, for example the pressure and quantities adsorbed.

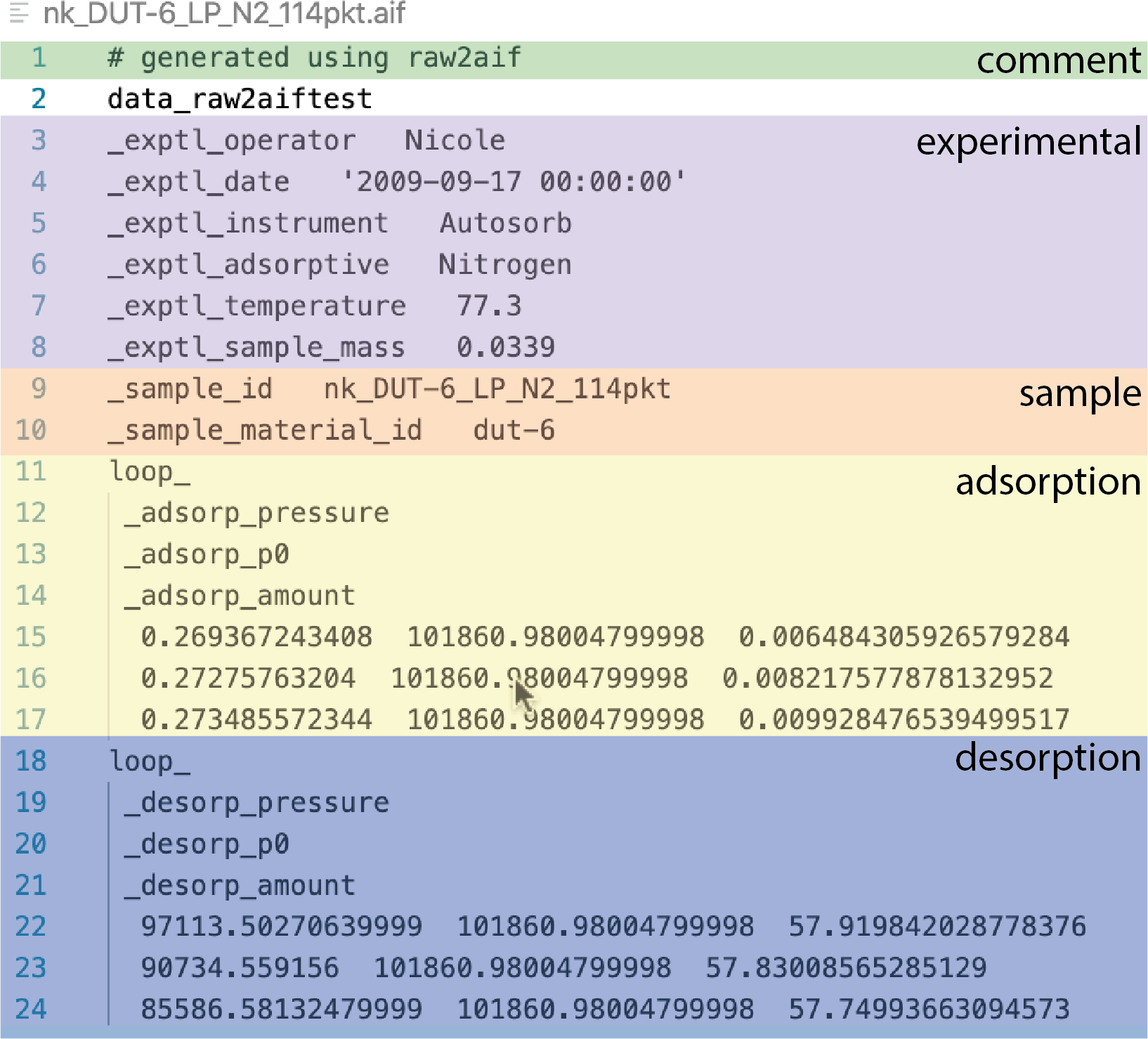


Figure 1: The structure of a typical adsorption information file (AIF) with the different data areas representing metadata (experimental and sample information) and the adsorption and desorption data.

95 The above data structures of the STAR format prove ideal for describing the detailed

96 quantities relating to an adsorption experiment. An entire experiment can be described

97 within a single data block. Initially, data name and data value pairs can be used to identify

98 quantities relating to the whole experiment such as the date and time, temperature and

99 adsorptive. The data points for the adsorption and desorption branch can then be detailed

100 in separate data loops. For each point, quantities such as the absolute pressure, relative

101 pressure and amount adsorbed can be detailed. This leads us to the general data structure

102 for the AIF (Figure 1).

103 **AIF dictionary**

104 Each data item within a data block is identified using a unique data name. The currently ac-

105 cepted AIF data names are listed and defined in Table 1. These data items form a first version

106 of a core dictionary, which are intended for primarily reporting primary physisorption data

107 for gas on porous materials. As a result the dictionary contains information about crucial ex-

108 perimental parameters, basic sample characterization, adsorption and desorption measures,

109 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, 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, mol kg*−*1)

110 The goal of the experimental data names is to capture the information required to ac-

111 curately portray the results of experiment, which at the very minimum are the temperature

112 and adsorptive. As metadata should include information that may aid in identifying possible

113 errors this also includes the mass of sample used in the experiment. Additionally, data is

114 included relative to the date, time and person responsible for the measurement to produce

115 a general picture of the provenance of the experiment.

116 Details of the sample are recorded for comparison and to uniquely identify the sample that

117 was used in the experiment. Simply, this includes the sample identity code ( sample id), a

118 laboratory code used to record and represent the sample and a string relating to the general

119 designation given to the material ( sample material name).

120 For each data point measured during an adsorption experiments data can be stored

121 that relates to the equilibrated pressure, the saturation pressure and the amount adsorbed.

122 Importantly, these should be stored in units of pascal and mol kg*−*1 to follow the IUPAC

123 recommendations.5 The saturation pressure can be recorded at each data point as we have

124 discovered that some instruments will calculate and report this value for every point.

125 The above features serve as an example of the data that can be archived in a AIF and can

126 be extended in future versions. For example, the intention of this initial version is to provide

127 digital archive of the presented primary data of an adsorption experiment, the isotherm. As a

128 result, it describes only basic information that usually exported from adsorption instruments

129 and a few convenient metadata items. The flexibility of this approach can be extended in

130 future versions to include more detailed parameters not commonly presented, such as dead

131 volume measurements, or combined with CIF data names to archive in situ measurements.

# 132 Creating and using the AIF

133 A simple python 3.7 routine was produced during the development of the AIF to convert

134 the plain text data output from Quantachrome and Belsorb instruments, used in our lab

135 and many others. Specifically, this has been written to parse raw analysis text files exported

136 by Quantachrome software and the raw data files (*.DAT* ) and CSV files exported by Bel-

137 sorb software. At this moment we are unable to read binary data files that are saved by

138 Quatachrome instruments (*.qps*).

139 This simple program will parse the output data to find data items detailed in the AIF

140 dictionary. These data items are converted to the appropriate units and the isotherm is split

141 into the adsorption and desorption parts. Subsequently an AIF is produced using *gemmi*

142 CIF routines.9 As a consequence routines and methods developed for treating CIF files can

143 be used in this application to adsorption data. For example produced AIF files can also be

144 read using existing CIF methods, such as the gemmi python library that is considered the the

145 fastest open-source CIF parser.9 We have used gemmi library to demonstrate how an isotherm

146 can be displayed from an AIF using a few lines. This is a very simple implementation and it

147 is hoped that future versions, following the cooperation of several manufactures, could result

148 in to routines to parse the information directly from binary data used by the instruments.

149 To illustrate the importance of a universal data format for adsorption data we can

150 compare data contained within an AIF to data digitized from a figure in a manuscript,

151 for the same experiment. The porous material DUT-6 (also referred to as MOF-206) is

152 a metal-organic framework produced from zinc nitrate and two different different ligands;

153 2,6-naphthalenedicarboxylic acid and benzene-1,3,5-tribenzoic acid. 10 DUT-6 has mesopores

154 with dodecahedral geometry and smaller cages that produce an extremely porous material

155 with large adsorption capacities for*n*-butane (up to 1.1 g g*−*1 at 293 K), methane (230 mg g*−*1

156 at 100 bar and 298 K) and hydrogen (60 mg g*−*1 at 50 bar and 77 K). The nitrogen adsorption

157 isotherm at 77 K was reported in the original manuscript using relative pressure and a linear

158 axis (Figure /reffgr:example). From this plot the isotherm can be classified as type Ib, clearly

159 exhibiting characteristics of a hierarchal pore system with micropores and mesopores. There

160 is evidently a distinct step *p/p*0 = 0.17, however, from this plot it is not possible to correctly

161 identify low pressure features. The low pressure regions is important for understanding the

162 process of initial pore filling or for independent application of analysis methods, such as

163 the Brunauer–Emmett–Teller (BET) method.11 The reported isotherm was digitized using

164 WebPlotDigitizer, which uses a time-consuming but powerful point-picking methodology.12

165 In this methodology, each point in the graph is identified by clicking on the image area

166 and the point’s location in on-screen pixels is then converted to data values, based on an

167 axes calibration. The digitized isotherm and the data from the experiment, when plotted

168 on semi-log axes, clearly shows important low pressure information, such as the point of ini-

169 tial pore-filling, is lost when this information is simply included as a figure in a manuscript

170 (Figure 2).

171 To prevent this information loss and to allow further analysis of reported adsorption

172 experiments by other researchers, adsorption data should be archived and shared using a

173 universal and complete data format like that of the AIF.

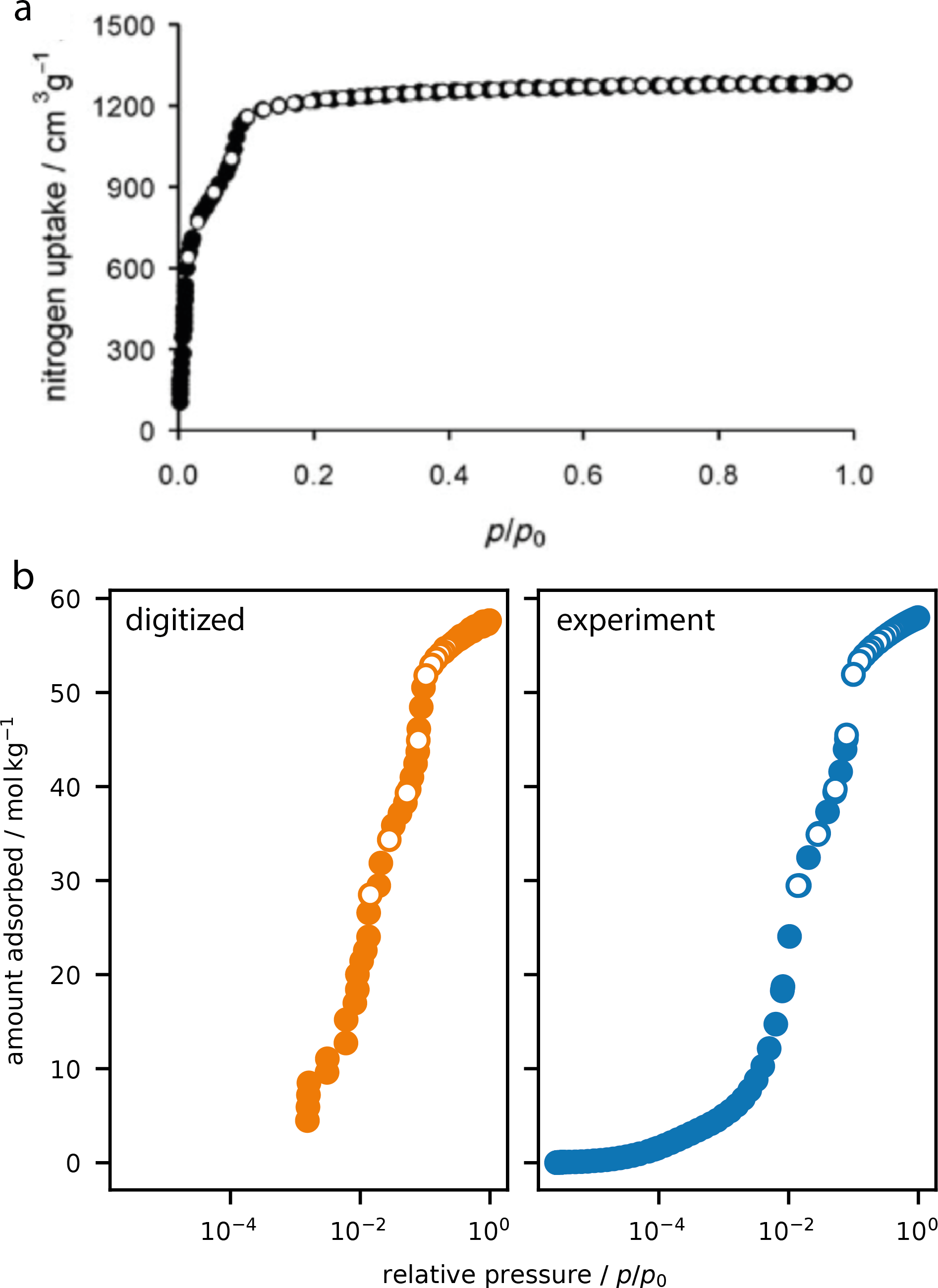


Figure 2: The reported nitrogen adsorption isotherm of DUT-6 recorded at 77 K (a) as reported in Ref. 10. A digitized representation of this reported isotherm was achieved using a point-picking method (b) and compared to the experimental data as contained within an AIF (c).

# 174 Summary and perspective

175 There is a current need to a universal format for archiving adsorption data. This is pertinent

176 for the development of detailed databases of information that can play an important role

177 in further materials design.13 This report details the specification of a new standard file,

178 named AIF, based on the self-defining text archive and retrieval (STAR) procedure, inspired

179 by similar universal and evolving formats currently used in the larger scientific community.

180 The AIF is importantly an easily extended free-format archive file that can be both human

181 and machine readable. This format represents the first steps to an open adsorption data

182 format and the community is encouraged to take up this format.

183 **Acknowledgement**

184 J. D. E. acknowledges the support of the Alexander von Humboldt foundation and HPC plat-

185 forms provided by a GENCI grant (A0070807069) and the Center for Information Services

186 and High Performance Computing (ZIH) at TU Dresden.

187 **References**

188 (1) Coudert, F.-X. Materials Databases: The Need for Open, Interoperable Databases

189 with Standardized Data and Rich Metadata. *Advanced Theory and Simulations* **2019**,

190 *2*, 1900131.

191 (2) Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management

192 and stewardship. *Scientific Data* **2016**, *3* .

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

194 813–821.

195 (4) Sun, M.-H.; Huang, S.-Z.; Chen, L.-H.; Li, Y.; Yang, X.-Y.; Yuan, Z.-Y.; Su, B.-L.

196 Applications of hierarchically structured porous materials from energy storage and con-

197 version, catalysis, photocatalysis, adsorption, separation, and sensing to biomedicine.

198 *Chemical Society Reviews* **2016**, *45*, 3479–3563.

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

200 Rouquerol, J.; Sing, K. S. Physisorption of gases, with special reference to the evaluation

201 of surface area and pore size distribution (IUPAC Technical Report). *Pure and Applied*

202 *Chemistry* **2015**, *87*, 1051–1069.

203 (6) Hall, S. R.; Spadaccini, N. The STAR File: detailed specifications. *Journal of Chemical*

204 *Information and Modeling* **1994**, *34*, 505–508.

205 (7) Hall, S. R.; Allen, F. H.; Brown, I. D. The crystallographic information file (CIF): a new

206 standard archive file for crystallography. *Acta Crystallographica Section A Foundations*

207 *of Crystallography* **1991**, *47*, 655–685.

208 (8) Ulrich, E. L.; Baskaran, K.; Dashti, H.; Ioannidis, Y. E.; Livny, M.; Romero, P. R.;

209 Maziuk, D.; Wedell, J. R.; Yao, H.; Eghbalnia, H. R.; Hoch, J. C.; Markley, J. L. NMR-

210 STAR: comprehensive ontology for representing, archiving and exchanging data from

211 nuclear magnetic resonance spectroscopic experiments. *Journal of Biomolecular NMR*

212 **2018**, *73*, 5–9.

213 (9) Global Phasing Ltd,; CCP4, GEMMI - library for structural biology. 2020; https:

214 //github.com/project-gemmi/gemmi.

215 (10) Klein, N.; Senkovska, I.; Gedrich, K.; Stoeck, U.; Henschel, A.; Mueller, U.; Kaskel, S. A

216 mesoporous metal-organic framework. *Angewandte Chemie International Edition* **2009**,

217 *48*, 9954–9957.

218 (11) Rouquerol, J.; Llewellyn, P.; Rouquerol, F. *Studies in Surface Science and Catalysis*;

219 Elsevier, 2007; pp 49–56.

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

221 (13)

# 222 Graphical TOC Entry

223

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: in- stead it will overflow the edge of the box.