The Science International Accord on Open Data in a Big Data World and the IUCr's response

Marvin L. Hackert[1], Brian McMahon[2], Luc Van Meervelt[3], John Helliwell[4]

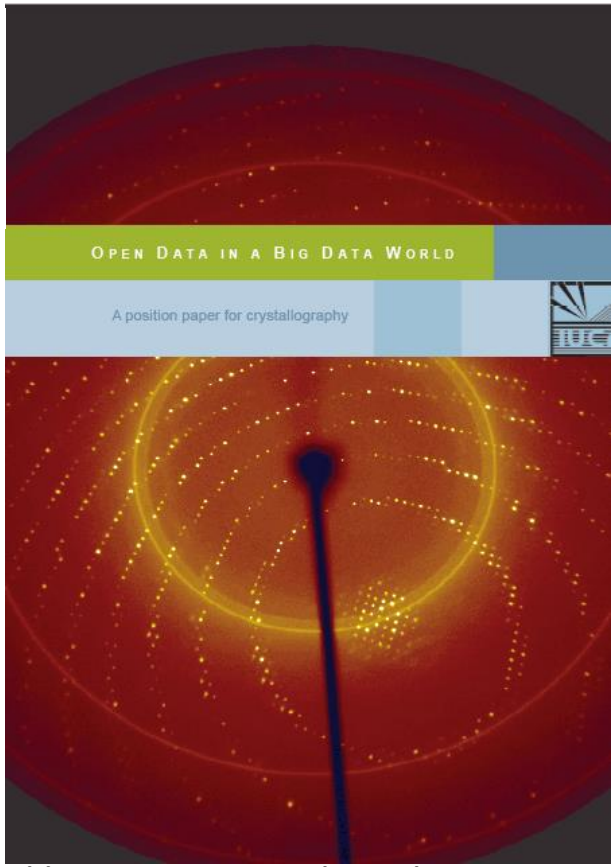
[1]IUCr President and IUCr Representative to ICSU
Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA

[2]IUCr Chester; [3]IUCr Sec-Treas., Leuven; [4]IUCr rep CODATA, Manchester

… encourages ***Open Data*** and ***Open Science*** practices

IUCr provided a detailed response:

https://www.icsu.org/publications/open-data-in-a-big-data-world



http://www.iucr.org/iucr/open-data

The IUCr and the crystallographic community has a tradition over many decades of open data linked to publication and supports the ICSU "Open Data in a Big Data World" report

The definition of "Open Data":

discoverable *– web search*
accessible *– easily imported*
intelligible *– relevant background data provided*
assessable *– assess issues on data quality*
*(re-)**usable** – metadata available to make data usable*

Endorsers

"Open Data in a Big Data World" accord passes 120 endorsements



INDIAN NATIONAL SCIENCE ACADEMY

THE NIGERIAN ACADEMY OF SCIENCE

editorial
Pontificia Universidad
JAVERIANA

UNION

NATURAL
HISTORY
MUSEUM


BANGLADESH ACADEMY OF SCIENCES



CLACSO

Presidencia de la Nación

ACADEMIA
BRASILEIRA
DE CIÊNCIAS

AKADEMIE DER WISSENSCHAFTEN

IUBS

ICSU
WORLD DATA SYSTEM



ZAGREB

НАЦІОНАЛЬНА
АКАДЕМІЯ
НАУК
УКРАЇНИ
1918

PACIFIC SCIENCE
ASSOCIATION

FaHCE

COSPAR


U.R.S.I.


COUNCIL OF FINNISH ACADEMIES



isprs

ANSALB

IdIHCS

iuto

SCIENCE COUNCIL OF JAPAN

ACAL
ACAL

futurearth
research for global sustainability

Understanding the U.S. Role in International Scientific Organizations:

ACA / AIP / USNCCr / IUCr / BISO / ICSU

ACA: **American Crystallographic Association** is a non-profit, scientific organization of ~ 1,300 members in more than 35 countries founded in 1949. The ACA is organized around **12 Scientific Interest Groups (SIGs).** The ACA is a member society of the **American Institute of Physics (AIP)** and a Regional Associate of the **International Union of Crystallography (IUCr)**.

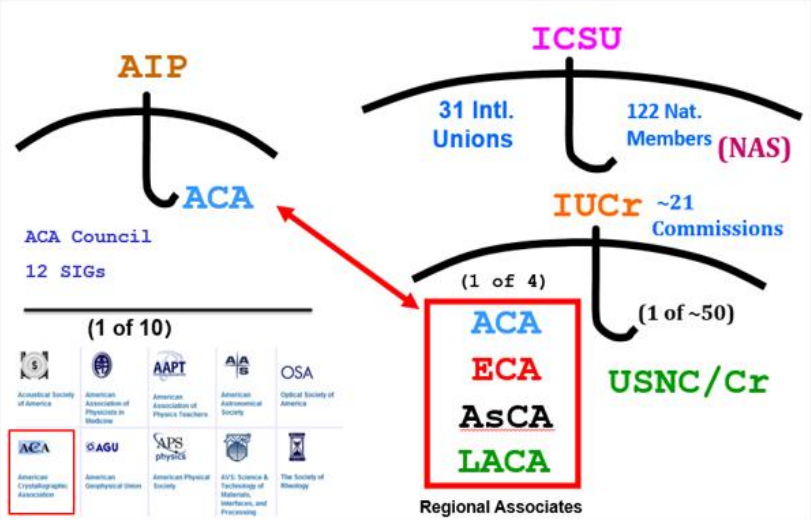
AIP: **American Institute of Physics** is a 501(c)(3) membership corporation of 10 physical science societies serving a combined membership of approximately 120,000 scientists, engineers and students. AIP publishes *Physics Today* .

USNC/Cr: The **U.S. National Committee for Crystallography (USNC/Cr)** represents U.S. crystallographers in the International Union of Crystallography **(IUCr)** and is supported by the National Academy of Sciences **(NAS).**

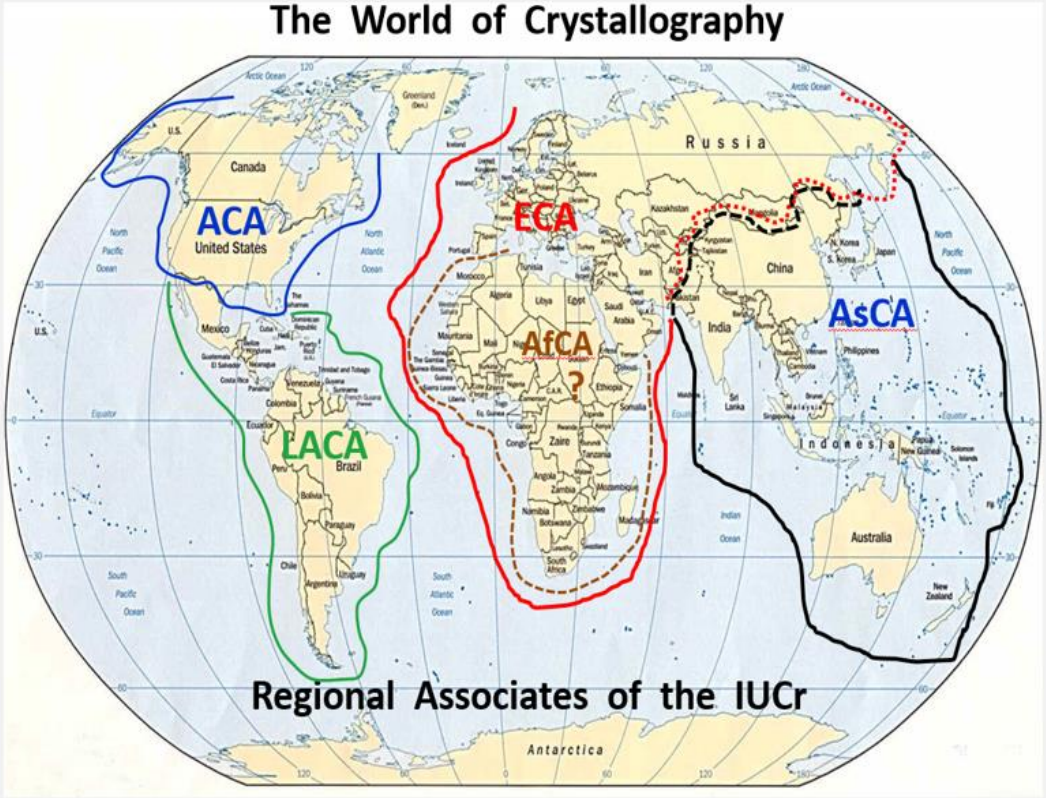
IUCr: The IUCr is an international scientific union that strives to promote international cooperation and publication in crystallography; to facilitate standardization of methods, units, nomenclatures and symbols; and to form a focus for the relations of crystallography to other sciences. The adhering bodies of the IUCr are the crystallographic national committees (USNC/Cr) of ~50 countries worldwide. The IUCr publishes **9 journals**, the **International Tables** of Crystallography, maintains an extensive **website** devoted to crystallography, **outreach** activities with crystallographic training, supports meetings around the world, and sponsors an **International Congress** every 3 years. The IUCr is organized around **21 Commissions**.

Role of Regional Associates: ACA / ECA / AsCA / LACA

AIP, IUCr & ICSU as umbrella scientific organizations



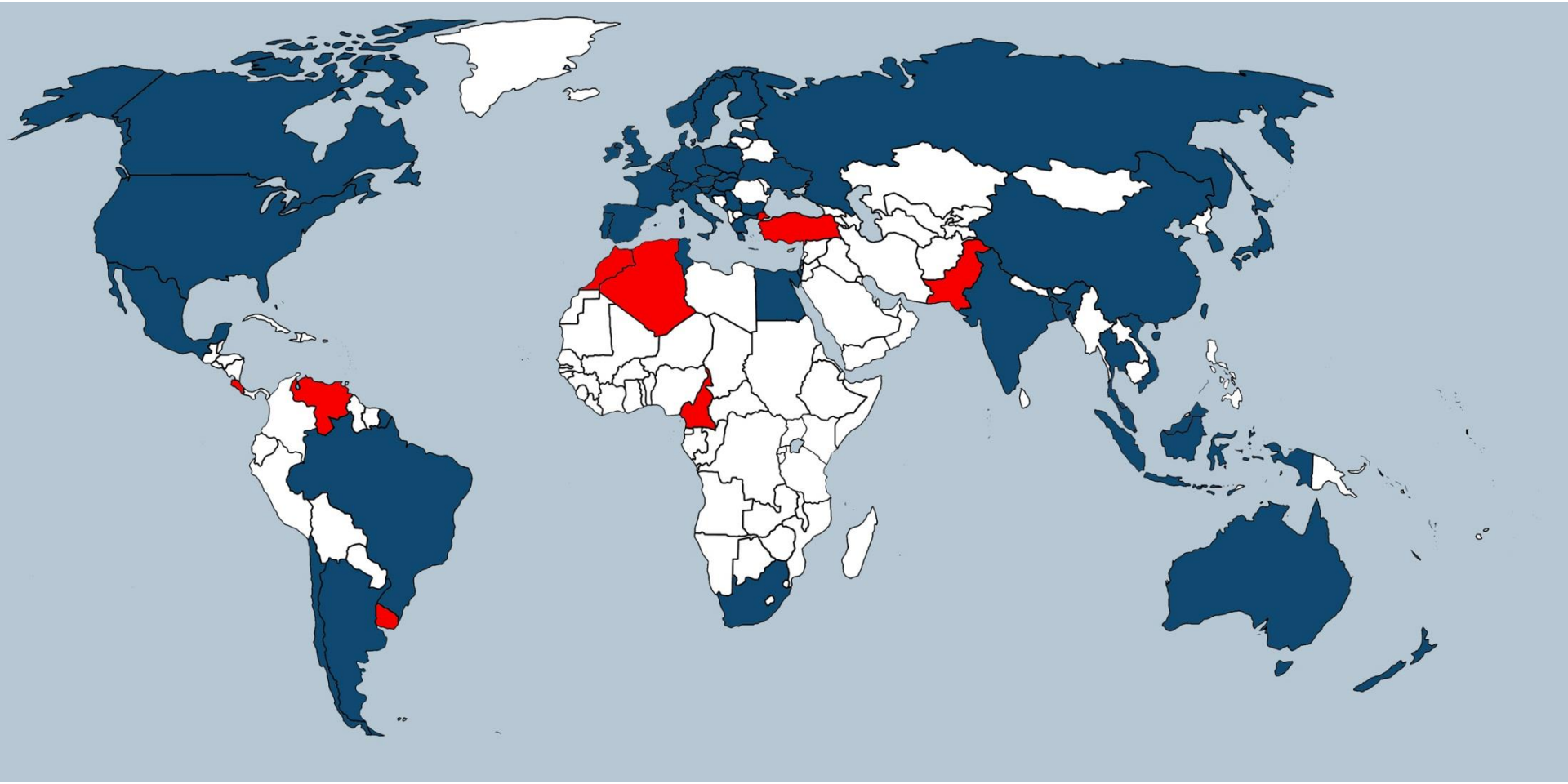
The ACA is a scientific society member of the AIP
The ACA is a Regional Associate of the IUCr



Asociación Latinoamericana de Cristalografía
LACA
Fundada en 2013

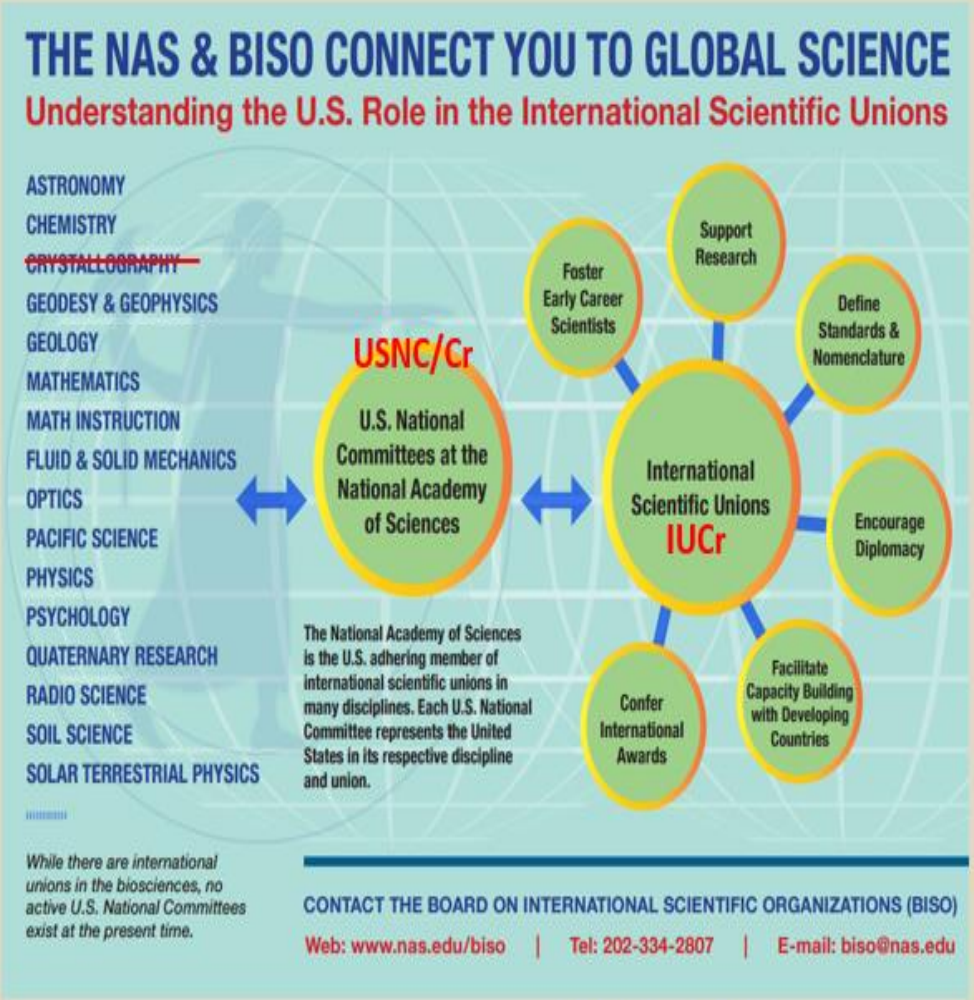
IUCr

Adhering Bodies of the IUCr, representing 52 countries, and its four Regional Associates



Countries that adhered to the IUCr before 2014 are shown in blue, those countries that joined in 2014 during IYCr are shown in red.

ICSU: The **International Council for Science (ICSU)** has **31 Scientific Union Members** (e.g. **IUCr**) and **122** National Scientific Members (e.g. **NAS**) covering **142** countries. Members provide input from a national, multidisciplinary perspective on priority areas for ICSU activities and facilitate linking national governments and science agencies. The majority of ICSU National Members are scientific academies. In the **U.S.**, the National Member is the **National Academy of Sciences**; in **Canada** it is the **National Research Council of Canada;** in the UK; it is the **Royal Society**.



National Sci. Members of ICSU.

Ref: Definitions and graphics taken from organization web sites.

Some history

The crystallographic databases are obviously very important: the early pioneers

J. D. Bernal ***Olga Kennard*** ***Walter Hamilton***

CSD (1965) > 875,000 **PDB (1971) > 130,000**

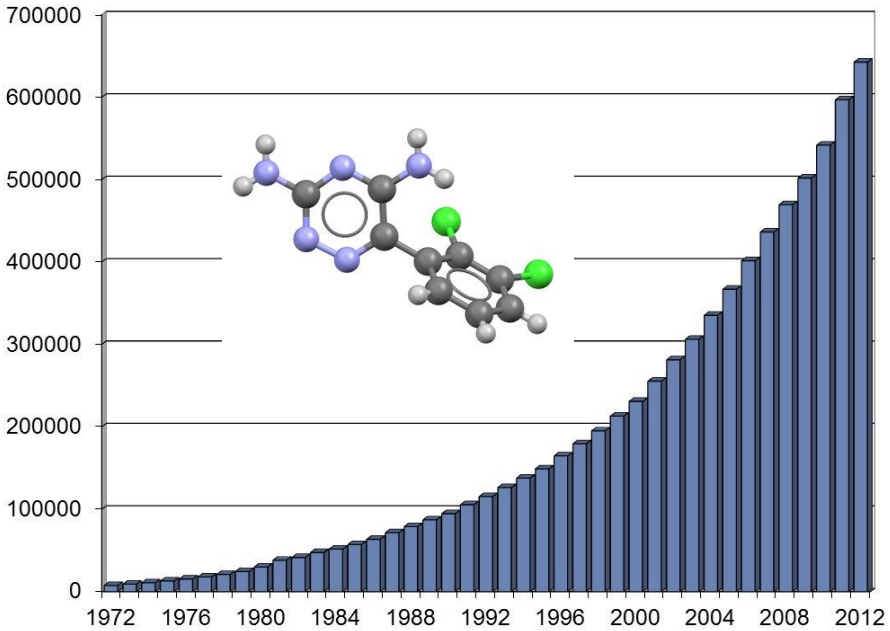
"Collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments"

Kennard, O. (1997). ***From Private Data to Public Knowledge.*** **In Butterworth, I. (Ed.)** ***The Impact of Electronic Publishing on the Academic Community. London:*** **Portland Press. pp. 159–166.**

… access to and re-use of other people's data is a legitimate route to new discovery

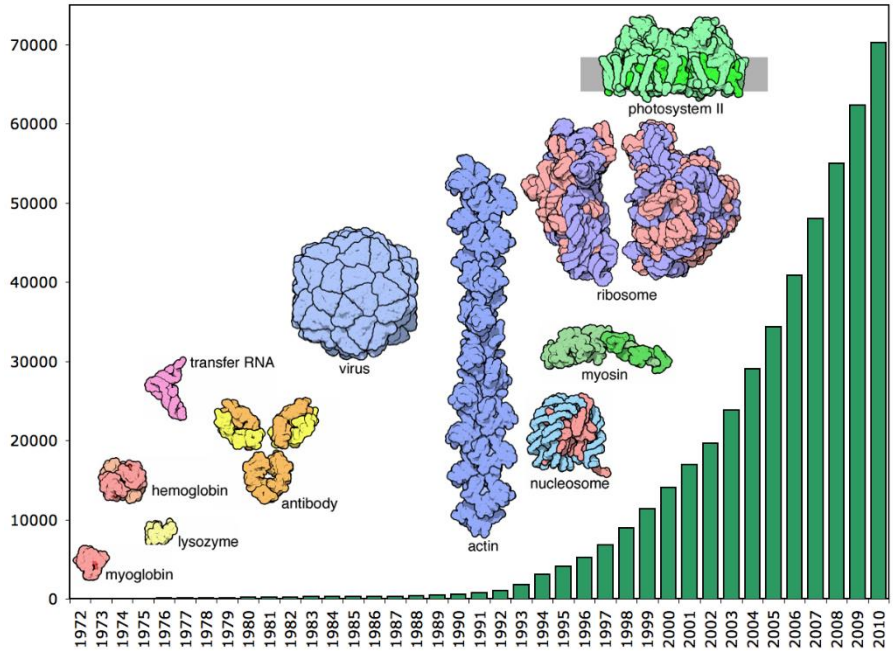
Structures large and small

CSD (1965) > 875,000



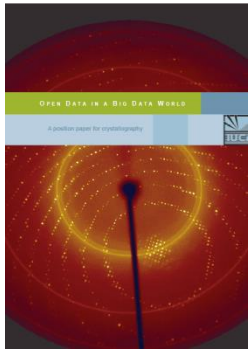
Cambridge Structural Database

PDB (1971) > 130,000



Protein Data Bank

Bruno, I., Helliwell, J. R., Gražulis, S., Kabekkodu, S., McMahon, B. & Westbrook, J. (2017). Crystallography and Databases. *Data Sci. J.* Submitted



In addition IUCr journals directly hold:

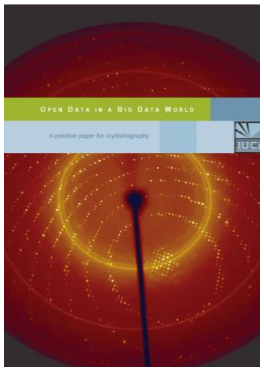
Chemistry and materials structural data sets, freely available from IUCr journals: > 58,800

Chemistry and Materials experimental intensity data sets (structure factors), freely available from IUCr journals: > 58,400

Chemistry and Materials experimental powder profile data sets, freely available from IUCr journals: > 1030

IUCr journals (open-access and hybrid) have a long history of data sharing – coordinates of derived structural models as well as underpinning experimental data (structure factors. etc.)

http://www.iucr.org/iucr/open-data



Openness alone is not sufficient:

Data validation and thereby data quality are needed

Data openly accessed ***must be subject to critical scrutiny****, through* ***peer review*** *and* ***automated validation*** *where possible*

The ***proper conduct of science has always depended*** *on a deep understanding of the nature of the data collected … and a* ***careful and proper analysis*** *of its accuracy, precision and validity*

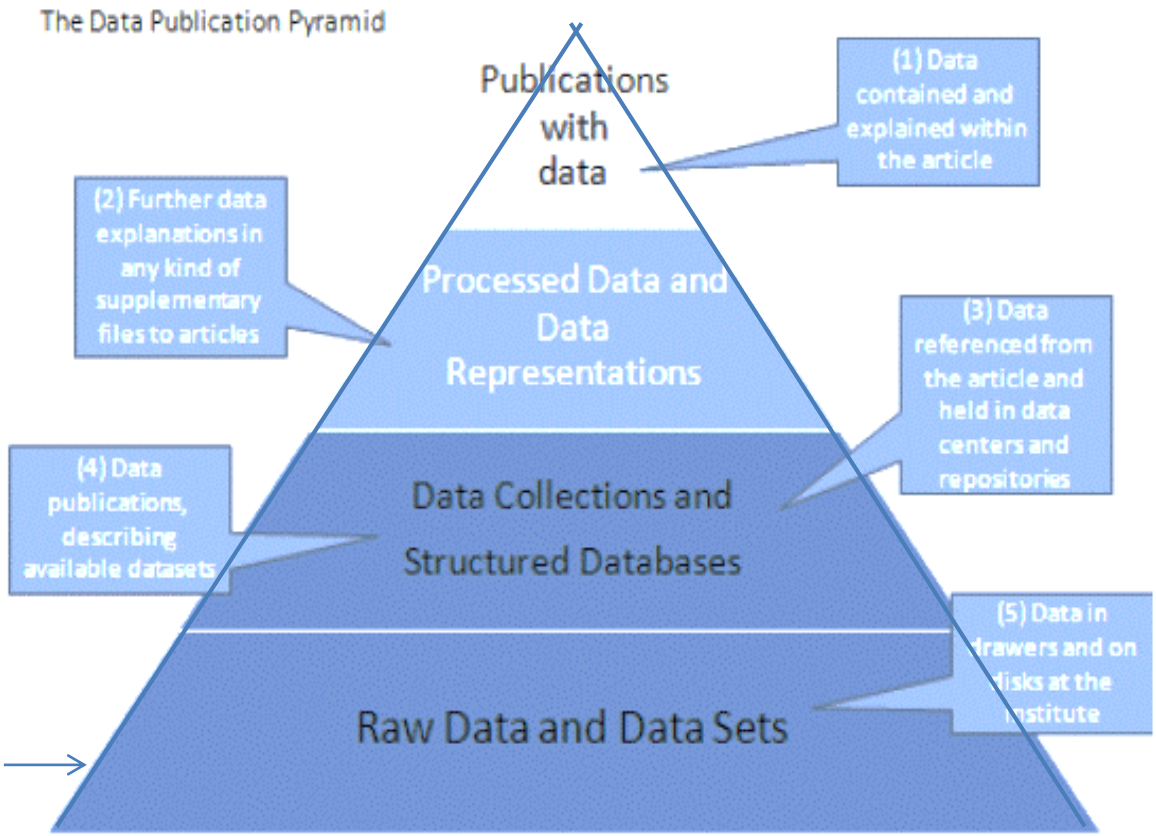
The ***IUCr formally documents every aspect of its data standardization*** *program on its website and through its journals and reference works*

There has been growing interest within the field of crystallography to ***retain the raw data*** *for each structure determination experiment*

→ ***new tools!***

http://www.iucr.org/iucr/open-data

The data publication pyramid shows a new challenge: raw data management!



***Big** data*

Reilly, S., Schallier, W., Schrimpf, S., Smit, E. & Wilkinson, W. (2011). *Report on Integration of Data and Publications.*

Available from http://www.stm-assoc.org/integration-of-data-and-publications/

IUCr Diffraction Data Deposition Working Group

To consider extending our community's best practice to raw data, in 2011 IUCr set up a ***Diffraction Data Deposition Working Group (DDDWG)***

http://www.iucr.org/resources/data/dddwg

Members of the DDDWG 2011 to 2017

- ***John R Helliwell and Brian McMahon (UK)*, Chair and Co-Chair**
- *Steve Androulakis (Australia)*
- *Sol Gruner (USA)/Doletha Szebenyi (USA)*
- *Loes Kroon-Batenburg (Netherlands)*
- *Tom Terwilliger (USA)*
- *John Westbrook (USA)*
- *Heinz-Josef Weyer (Switzerland)* †
- *Edgar Weckert (Germany)*

Recommendations from the DDDWG for the Triennium 2014 to 2017

- IUCr Commissions to define their ***metadata***;
- *J. Appl. Cryst.* to introduce a 'Difficult Raw Data' Section (Loes Kroon-Batenburg);
- A centralized crystallographic repository of raw data set metadata should be scoped, including a search interface, leading to a pilot service;
- With a viable pilot metadata registry ***authors should*** provide a ***permanent and prominent link*** from an article to their raw data sets underpinning a journal publication.

digital object identifier

A **digital object identifier** (DOI) is a unique alphanumeric string assigned by a registration agency (the International DOI Foundation) to identify content and provide a persistent link to its location on the Internet. The publisher assigns a DOI when your article is published and made available electronically.

en.wikipedia.org

IUCr applauds significant pioneering developments in preserving raw diffraction data

- Australian synchrotron MX raw data archive https://store.synchrotron.org.au/public_data/
- The USA NIH funded various structural genomics projects with raw data archives
- The Institut Laue Langevin and ISIS are exemplar at preserving all data and with DOIs (Digital Object Identifiers), *e.g.*

Science & Technology Facilities Council

ISIS Data Journal

The archive for ISIS research data

ISIS | ISIS Data

RB920486

Investigation title: Electric field effect on the interfacial uncompensated spins in the Co/BiFeO3/STO exchange bias system.

Release date: Fri Jul 26 09:06:29 BST 2013

Creator: *Dr Nina-Juliane Steinke*

DOI: 10.5286/ISIS.E.24079627

Date of Experiment: Fri Jul 23 08:52:43 BST 2010

Publisher: STFC ISIS Facility

Data format: RAW/Nexus
Select the data format above to find out more about it.

Data Citation

The recommended format for citing this dataset in a research publication is as:
[author], [date], [title], [publisher], [doi]

DOWNLOAD
download the dataset

Data collected
CRISP ins
at the ISIS

IUCr applauds further recent developments

additional repositories / uniform metadata standards

- University data repositories with ***DOI registrations***.
- EU's ***Zenodo*** science data archive (free of charge!) Open Science knows no borders!

  Fair Principles: findable / accessible / interoperable / reusable
- The ***Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC),*** led by Wladek Minor (U. Virginia, USA) http://www.proteindiffraction.org [Grabowski *et al.* (2016), *Acta Cryst.* D**72**, 1181-1193]
- The ***PDB*** now requests the information on raw data and metadata for raw data during a deposition *i.e.* their ***DOIs***
- ***ESRF Data Archive*** ("every raw data set measured with ***a registered DOI")***
- **IUCrData** (initially, *derived* data sets)
- The ***Structural Biology Data Grid*** has been launched [Meyer *et al.* (2016), *Nature Commun.* **7**:10882]

digital object identifier

A **digital object identifier** (DOI) is a unique alphanumeric string assigned by a registration agency (the International DOI Foundation) to identify content and provide a persistent link to its location on the Internet. The publisher assigns a DOI when your article is published and made available electronically.

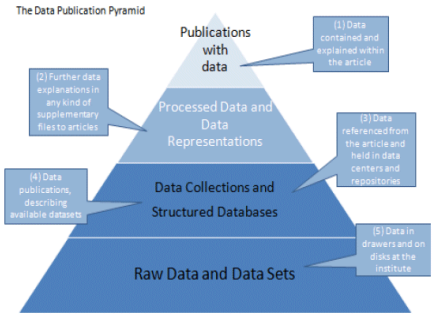
EUROPEAN SYNCHROTRON RADIATION FACILITY

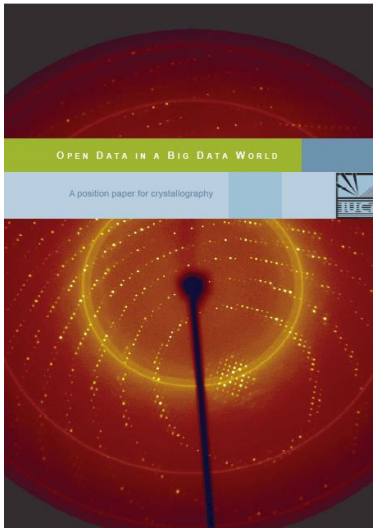
The ESRF Data Policy

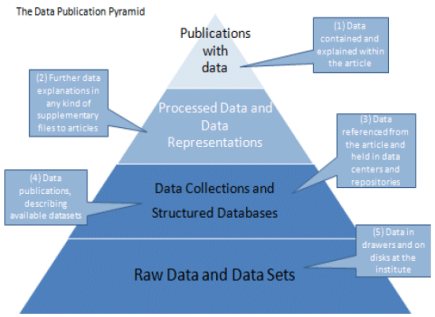
The ESRF aims to implement a Data Policy starting as soon as possible in 2016. The main elements of this policy comprise:

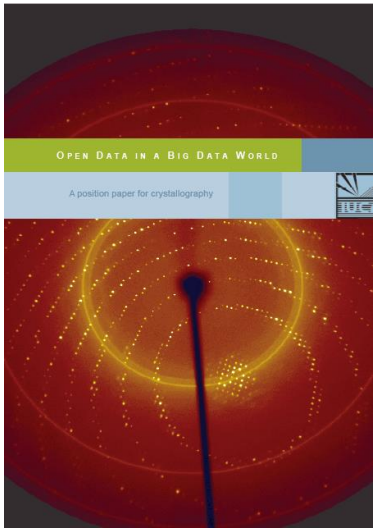
- **Data ownership**
- **Data curation**
- **Data archiving**
- **Open access to data**

This policy follows largely the recommendations of the PaN-data Europe Strategic Working Group laying out a common framework for scientific data management at photon and neutron facilities (Deliverable D2.1, PaN-data Europe, co-funded by the European Commission under the 7th Framework Programme)









Thank you
and
Enjoy the Workshop here at
ACA 2017
New Orleans