

Diffraction Data Deposition and Publication

Kay Diederichs, Manfred Weiss

Vienna, 18/08/2019

Workshop on Data Science Skills in Publishing: for authors, editors and referees

Organized by
IUCr Committee on Data



Sunday August 18 2019

Vienna, Austria

There is a trend towards ensuring that modern science research data are findable, accessible, interoperable and reusable (FAIR). However, this is something that crystallographers have been achieving for many decades, during which excellent crystallographic databases have always exploited the best available hardware for digital archiving. FAIR is necessary but not sufficient, as physicists would say, as the archived data should also be true facts. So FACT and FAIR are needed for reproducibility. The crystallographic community has developed automatic checking software by pooling its experiences from hundreds of thousands of crystal structure analyses into validation procedures with numerous data file checks on both coordinates and processed diffraction data sets. Alarm alerts can then be scrutinised by journal editors and referees. With such exemplary procedures is there anything to be improved? Crystallographers conclude that there is. Firstly the IUCr journal *Acta Cryst. C: Structural Chemistry* has always required submission of article with validation report with underpinning data files. Thus the specialist subject expertise of referees can involve their own direct calculations to supplement the automatic checks before article and data set acceptance as versions of record by the editor. This has inspired others to look to improve their own crystallographic disciplines and journals to follow the *Acta Cryst. C* standard. Secondly the digital archives have enhanced their capacity in recent years owing to amazing hardware advances so that even the Gigabyte-sized raw data sets can also be preserved as versions of record. A reader of a publication can thereby revisit even the earliest calculation decisions of the authors of a publication. As the Royal Society of London puts it: science is about not taking someone's word and so, instead, the science is always in the data. FACT and FAIR, indeed scientific objectivity itself, is possible. This Workshop will address the state of the art in the field and the data science skills hoped for, indeed to be expected, of all those involved in publishing crystallography results, and of results from all the cognate methods such as scattering, microscopy and spectroscopy.

Logistical information  (4.4 MB)

Reminder: The CommDat user forum is at <https://forums.iucr.org>

“modern science research data being Findable, Accessible, Interoperable and Reusable i.e. FAIR”

Requirements for publication in (serious) scientific journals:

- Macromolecular coordinates: mandatory deposition in PDB (created 1971)
- Structure factors: mandatory to accompany coordinates since 2008

“FAIR is necessary but not sufficient, as physicists would say, since the archived data should also be true facts”

- archived data should not only allow to reproduce the results of the crystallographic experiment, in particular the maps and coordinates (which one CAN in principle do with deposited structure factors)
- but the archived data should also allow to identify and correct wrong procedures and interpretations made by the original authors: needs the raw data
- these provide – to a large extent - safeguards against untrue, namely made-up (faked) data.

**“As the Royal Society of London puts it:
science is about not taking someone’s word
and so, instead,
the science is always in the data”**

- the data (raw measurements plus metadata) must be provided/available to allow for unbiased and independent processing and interpretation.
- unmerged intensities do not provide the level of truth that raw data guarantee, because they depend on the skills of the authors, the data processing programs used, and their options.
- no standard exists for the specification of “unmerged data” – should they be scaled or unscaled, which items are important, what about serial crystallography, ...

IUCr Diffraction Data Deposition Working Group

- ... has been investigating (since 2011) the *rationale and policies for routine deposition of diffraction images (and other primary experimental data sets)*.
- “Federated repositories of X-ray diffraction images”, Androulakis et al. (2008) **Acta D64**, 810
- “Experiences with archived raw diffraction images data: capturing cisplatin after chemical conversion of carboplatin in high salt conditions for a protein crystal”. Tanley, Diederichs, Kroon-Batenburg, Schreurs, Helliwell (2013) **J. Synchrotron Rad.** **20**, 880
- “Archiving raw crystallographic data”, T. Terwilliger (2014) **Acta D70**, 2500
- “Experiences with making diffraction image data available: what metadata do we need to archive?”, Kroon-Batenburg & Helliwell (2014), **Acta D70**, 2502
- “How to make deposition of images a reality”, Guss & McMahon (2014) **Acta D70**, 2520
- “Continuous mutual improvement of macromolecular structure models in the PDB and of X-ray crystallographic software: the dual role of deposited experimental data”, Terwilliger & Bricogne (2014) **Acta D70**, 2533
- “Findable Accessible Interoperable Re-usable (FAIR) diffraction data are coming to protein crystallography”, Helliwell, Minor, Weiss, Garman, Read, Newman, v.Raaij, Hajdu, Baker (2019) **Acta D75**, 455 (same editorial in **J. Appl. Cryst.**, **IUCrJ**, **Acta F**)

“A reader of a publication can thereby revisit even the earliest calculation decisions of the authors of a publication.”

This affects:

- | | |
|---|---------------------|
| – Authors | Short-term benefit |
| – Editors | “ |
| – Reviewers | “ |
| – Readers of the final published manuscript | “ |
| – End user of the PDB entry (biology, medicine, pharmacology, ...) | “ |
| – Software developers | Medium-term benefit |
| – Researchers (investigate features of data like diffuse scattering, ...) | Long-term benefit |
| – <u>Community / Science</u> | “ |

in many different ways!

Authoring, editing, refereeing a paper – what changes?

- Authors upload raw data set(s) and metadata to SGrid, Zenodo, proteindiffraction.org, ... and document its DOI in Table 1
 - Speaking of Table 1 ... does it fulfill the needs?
 - “Against Method: Table 1 – Cui Bono?”, Rupp (2018) **Structure** 26, 919 : Table 1 *“is a relic from pre-structure factor deposition times, does not inform about structure model quality, contains data items of limited or disputed usefulness, includes data items that do not inform a non-specialist reader, includes data items frequently not correctly interpreted, is largely redundant but often inconsistent with the PDB record of entry”*
 - Different people expect different items in Table 1, but availability of raw data would allow to dynamically extract or reproduce those items that are not, or cannot be represented (e.g. 3D StarAniso visualizations)
 - Editors and referees of scientific journals may consult the raw data when in doubt, but this is not anticipated to happen often UNLESS web services are established that automate reliable and standardized data processing!
 - Raw data archival is just the start; what is then needed are
 - * consistent requirements across journals
 - * making the added value of raw data easily accessible
-

Using the archived data after publication – what changes?

- Readers and users may critically assess the procedures and decisions

Recent examples for “crowd analysis” on CCP4BB:

- “Questionable Ligand Density: 6MO0, 6MO1, 6MO2”, Rhys Grinter, July 19; 29 postings analyzing the deposited data (*not* the raw data!), and culminating in the identification of a number of further questionable ligand papers (JBC 2013, PNAS 2012, Sci.Rep. 2017) – see <https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=CCP4BB;446d0f0d.1907>
- “[6HR5] collected on an Eiger so Rmerge not relevant”, Weston Lane, July 31; 12 postings – summary by C. Vonrhein: *“Rfree of 36% seems really high ... If you look at the maps (e.g. after some re-refinement with your favourite refinement package) it seems as if there are a few sequence shifts (around and after A78), some poor density and additional unmodelled density ... which all add to those high R-values I guess. But given the poor data quality and no raw images (to check and maybe improve upon that), I didn't feel the urge to delve into that any further ;-”*
<https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=CCP4BB;35d2606e.1908>
- Re-processing is warranted if validation statistics (traffic lights!) are poor
- With raw data archival in place, I expect re-processing services similar to PDB-REDO

Automated (re-)processing

“Automation is a good servant, but a poor master”

Current pipelines give good results for (up to) 90% of synchrotron data sets.

But the remaining >10% are critical.

Many - if not most - users

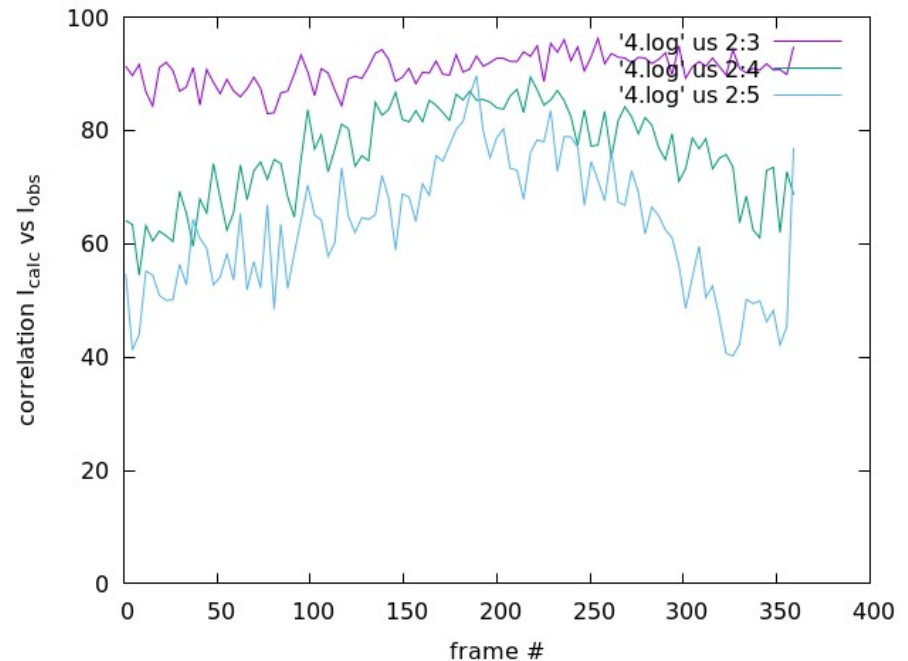
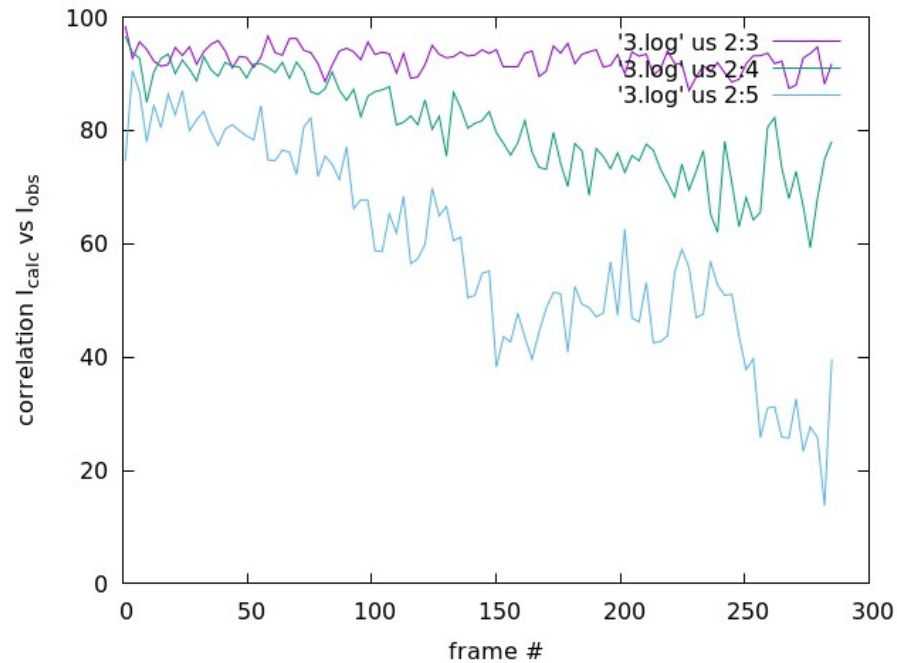
- have little or no crystallographic experience/knowledge
- are frightened by space group determination, twinning, ...
- do not inspect their raw data visually, nor the control images from processing
- rely on existing software to produce Table 1, to formally satisfy editors/referees/readers

Automatic (re-) processing software must answer the following questions reliably:

- are the raw data processed properly (resolution limits, shadows)?
- has the Bravais lattice / space group been assigned properly?
- have weak reflections been missed (under-prediction)?
- are too many weak reflections in the data, indicating a smaller cell (over-prediction)?
- have other peculiarities (eg twinning) been missed?
- do the data suffer from radiation damage, and how much?

New tools for data (re-)processing:

1) compare I_{obs} with I_{calc}



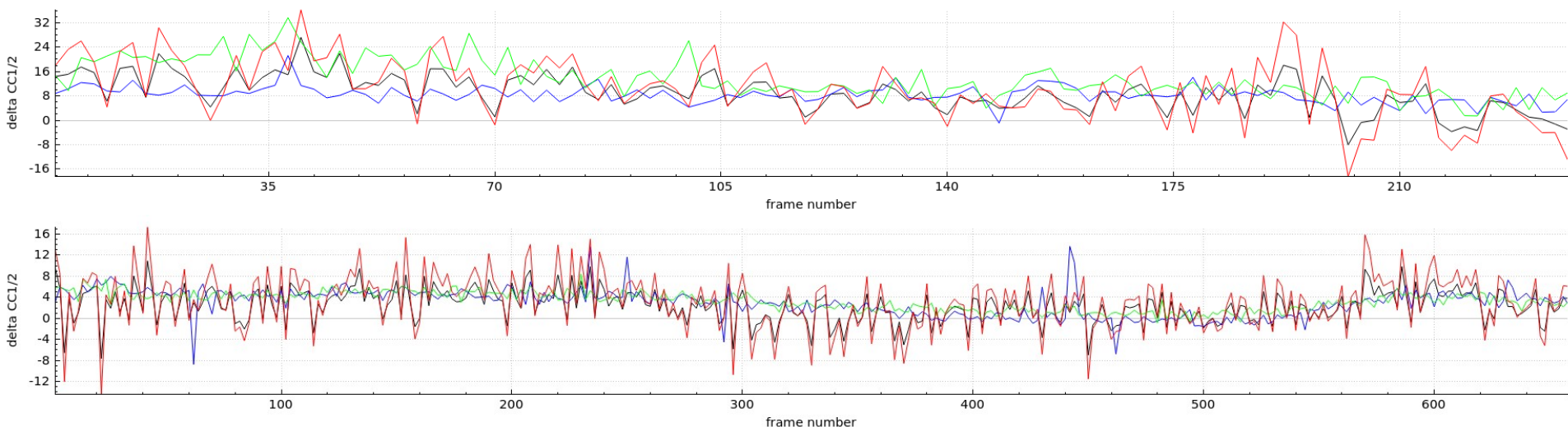
- Identify bad data sets and data wedges; assess radiation damage
- Iterative procedure – need to re-refine to get updated I_{calc}

New tools for data (re-)processing:

2) The $\Delta CC_{1/2}$ method

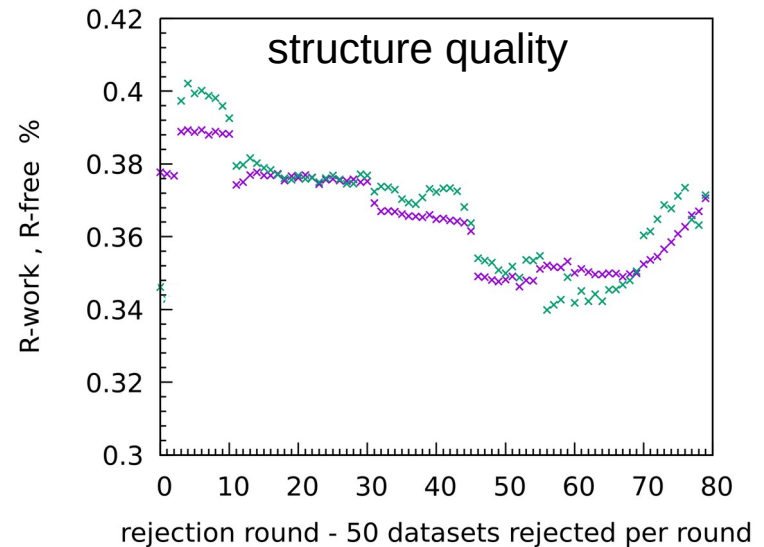
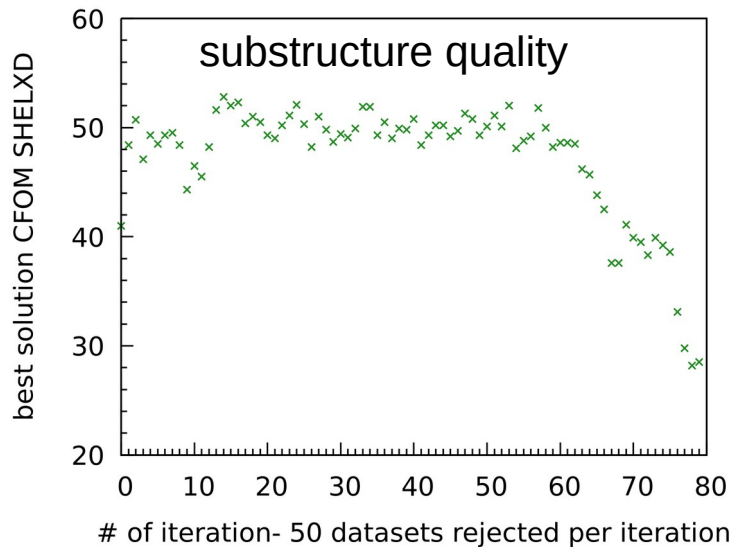
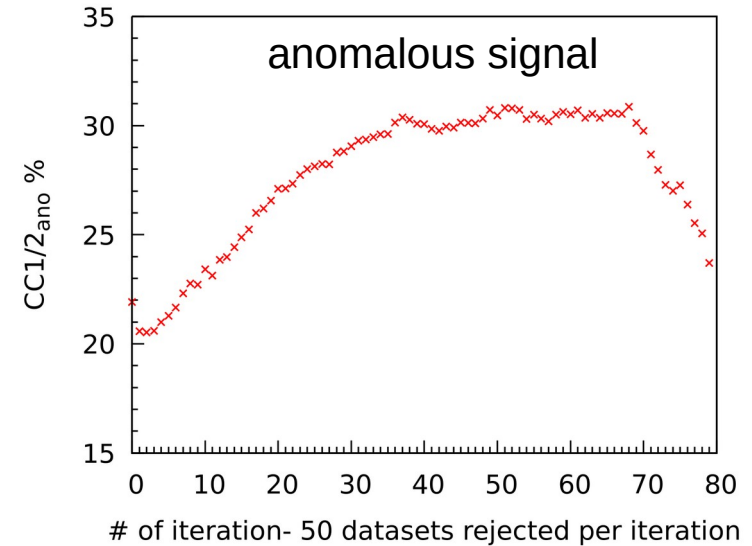
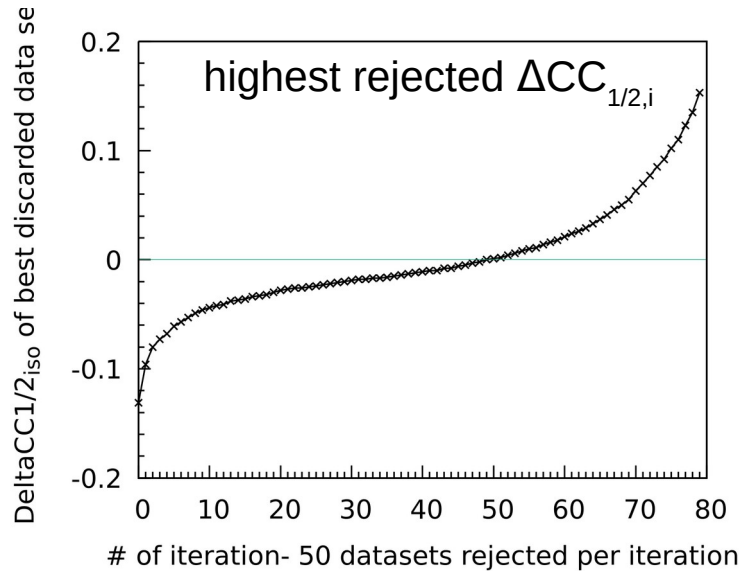
(Assmann, Brehm and Diederichs (2016) J. Appl. Cryst. 49, 1021-1028); XDSCC12

- $\Delta CC_{1/2,i} = CC_{1/2,\text{all}} - CC_{1/2,\text{all_except_i}}$
- $CC_{1/2} = \frac{\sigma_I^2}{\sigma_I^2 + \sigma_e^2} = \frac{\sigma_y^2 - \frac{1}{2}\sigma_e^2}{\sigma_y^2 + \frac{1}{2}\sigma_e^2}$
- Single data set: find bad parts; radiation damage
- Multiple data sets: discard bad data sets



Example: 4G4A (HEWL + cisplatin @ RT) data sets 3 and 9

Serial crystallography: discard data sets with most negative $\Delta CC_{1/2,i}$



This validates the choice of $CC_{1/2}$ as optimization target !

Summary

- Deposition of raw diffraction data is possible and practical now!
- To be practically useful on a large scale, requires servers/software for automated (re-)processing
- New software tools are available to extract additional information, beyond Table 1
- This added information improves processing and the resulting structures, and informs the biological understanding

Workshop on Data Science Skills in Publishing: for authors, editors and referees

Organized by

IUCr Committee on Data



Sunday August 18 2019

Vienna, Austria

There is a trend towards ensuring that modern science research data are findable, accessible, interoperable and reusable (FAIR). However, this is something that crystallographers have been achieving for many decades, during which excellent crystallographic databases have always exploited the best available hardware for digital archiving. FAIR is necessary but not sufficient, as physicists would say, as the archived data should also be true facts. So FACT and FAIR are needed for reproducibility. The crystallographic community has developed automatic checking software by pooling its experiences from hundreds of thousands of crystal structure analyses into validation procedures with numerous data file checks on both coordinates and processed diffraction data sets. Alarm alerts can then be scrutinised by journal editors and referees. With such exemplary procedures is there anything to be improved? Crystallographers conclude that there is. Firstly the IUCr journal *Acta Cryst. C: Structural Chemistry* has always required submission of article with validation report with underpinning data files. Thus the specialist subject expertise of referees can involve their own direct calculations to supplement the automatic checks before article and data set acceptance as versions of record by the editor. This has inspired others to look to improve their own crystallographic disciplines and journals to follow the *Acta Cryst. C* standard. Secondly the digital archives have enhanced their capacity in recent years owing to amazing hardware advances so that even the Gigabyte-sized raw data sets can also be preserved as versions of record. A reader of a publication can thereby revisit even the earliest calculation decisions of the authors of a publication. As the Royal Society of London puts it: science is about not taking someone's word and so, instead, the science is always in the data. FACT and FAIR, indeed scientific objectivity itself, is possible. This Workshop will address the state of the art in the field and the data science skills hoped for, indeed to be expected, of all those involved in publishing crystallography results, and of results from all the cognate methods such as scattering, microscopy and spectroscopy.

Logistical information  (4.4 MB)

Reminder: The CommDat user forum is at <https://forums.iucr.org>

IUCr forums

Discussions on IUCr projects and activities

FAQ Search

[Board index](#) / [Standing Committees and Working Groups](#) / [Public input to CommDat](#)















Public input to CommDat

[New Topic](#)

Search this forum...



18 topics • Page 1 of 1

TOPICS	STATISTICS	LAST POST
 Lecture at the ACA 2019 Transactions Symposium on Data by JRH » Tue Jul 30, 2019 10:20 am	Replies: 0 Views: 692	by JRH  Tue Jul 30, 2019 10:20 am
 Reports describing the work of the IUCr CommDat and various Commissions by JRH » Tue Jul 02, 2019 4:19 pm	Replies: 0 Views: 29	by JRH  Tue Jul 02, 2019 4:19 pm
 Abstracts for the ECM32 Satellite on Data Science Skills in Publishing for Editors, Authors and Referees by JRH » Wed Jun 26, 2019 8:16 am	Replies: 0 Views: 320	by JRH  Wed Jun 26, 2019 8:16 am
 FAIR diffraction data and IUCr journals by Brian McMahon » Thu May 09, 2019 5:52 pm	Replies: 0 Views: 596	by Brian McMahon  Thu May 09, 2019 5:52 pm
 The UK Open Research Data Task Force Final Report by JRH » Mon Mar 18, 2019 4:40 pm	Replies: 0 Views: 288	by JRH  Mon Mar 18, 2019 4:40 pm
 Crystallographic Information Fiesta! by Brian McMahon » Thu Feb 14, 2019 12:02 pm	Replies: 0 Views: 309	by Brian McMahon  Thu Feb 14, 2019 12:02 pm
 IUCr Workshop on Data Science Skills in Publishing by JRH » Fri Jan 18, 2019 7:38 pm	Replies: 0 Views: 332	by JRH  Fri Jan 18, 2019 7:38 pm

**Thank you
for your interest!**

Request PDF of this talk from Kay.Diederichs@uni-konstanz.de