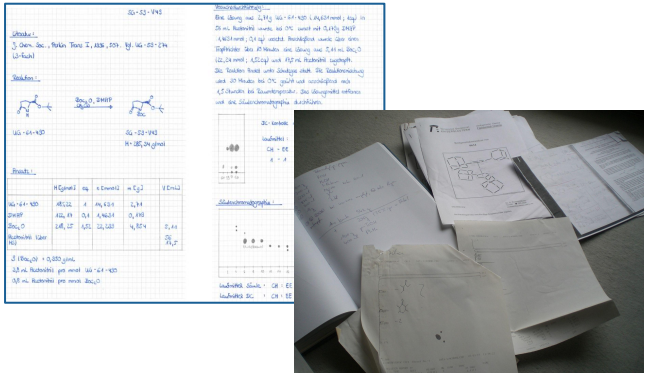
Slides: https://t1p.de/r1b6w


# NFDI4Chem: Auf dem Weg zu FAIRen NMR-Daten

Tillmann Fischer, Computergestützte Pflanzenbiochemie, Leibniz-Institute für Pflanzenbiochemie, Halle (Saale), PPNMR 2023, 21. März 2023, 14:30 s. t., Max Delbrück Center, Berlin/Buch

# Der Status Quo

Analog / Digital

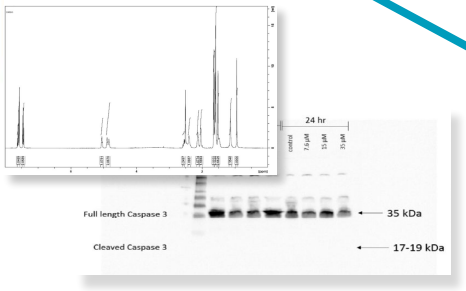


Digital



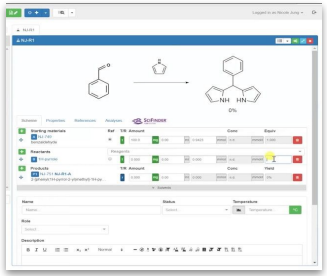







nicht maschinenlesbar

# NFDI4Chem – NFDI Konsortium für Chemie

Elektronische Laborjournale

# FAIR ist kein Modewort, es ist eine Einstellung!

**F**indable

**A**ccessible

**I**nteroperable

**R**eusable

- **Findable**: Leicht auffindbar, unter Verwendung eines eindeutigen und **dauerhaften Identifikatoren** (PIDs). **Beschreibende Metadaten**, die das Auffinden ermöglichen.
- **Accessible**: Leichtes Abrufen durch Maschinen und Menschen in **Repositorien** unter Verwendung von **Standardprotokollen**, mit Metadaten und Archivierung.
- **Interoperable**: Verfügbar in **Formaten**, die ausgetauscht und auch nach langer Zeit interpretiert werden können, einschließlich Metadaten.
- **Reusable**: **Metadaten** gewährleisten die Wiederverwendung für zukünftige Forschung,mit klarer **Lizenz** und **Provenienz**, entsprechend den **Gemeinschaftsstandards**.

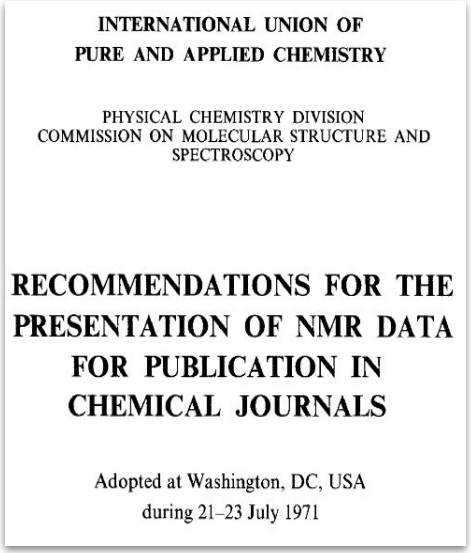
Adaptiert aus: The FAIR Data Principles for Research Data, URL: https://blogs.tib.eu/wp/tib/2017/09/12/the-fair-data-principles-for-research-data/ .

# Herausforderungen

Lücken im Datenlebenszyklus von NMR-Daten:

**Publikation von NMR-Daten (NMR Text)**
-> Basierend auf IUPAC Empfehlungen von 1971 und 1975. Messdaten werde mehrheitlich nicht publiziert.

**Wiederverwendung von NMR-Daten**
-> schwierig, da Daten oft nicht interoperabel.



INTERNATIONAL UNION OF
PURE AND APPLIED CHEMISTRY

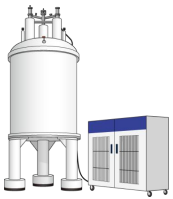
PHYSICAL CHEMISTRY DIVISION
COMMISSION ON MOLECULAR STRUCTURE AND SPECTROSCOPY

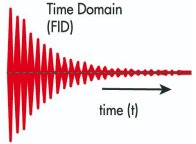
RECOMMENDATIONS FOR THE PRESENTATION OF NMR DATA FOR PUBLICATION IN CHEMICAL JOURNALS

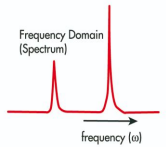
Adopted at Washington, DC, USA
during 21–23 July 1971

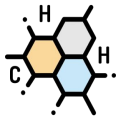
*Pure Appl. Chem.* **1972**, *29,* 625, DOI: 10.1351/pac197229040625. *Pure Appl. Chem.* / IUPAC Standards Online **1976**, *45,* 217, DOI: 10.1515/iupac.45.0031.

# Quellen von NMR Dateien & Workflows






RAW FILES

JOEL .jdf

BRUKER .zip

ACD/Labs

NMR NOVA

NMR SOFTWARE TopSpin

NMRium

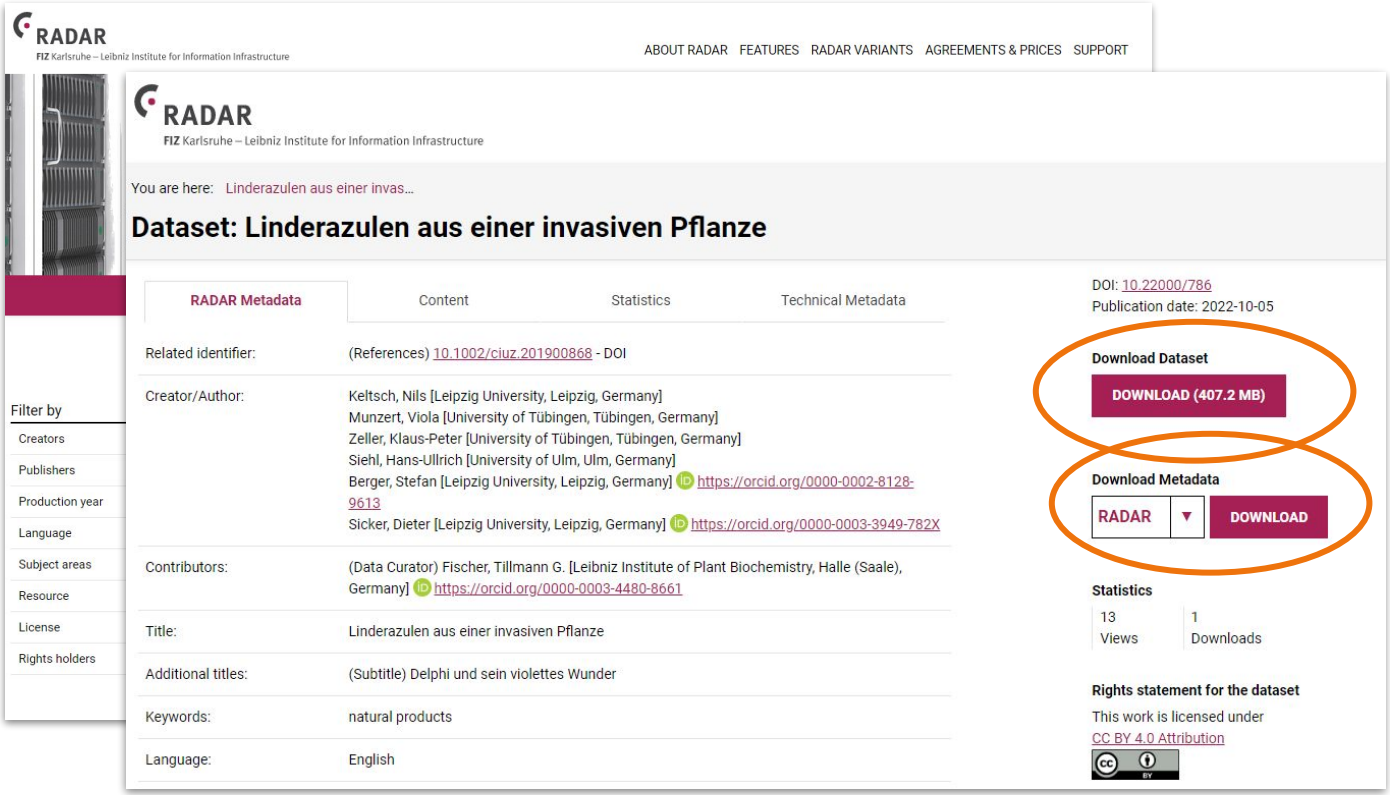
PROCESSED FILES ~ STANDARD FORMATS

nmreData .sdf

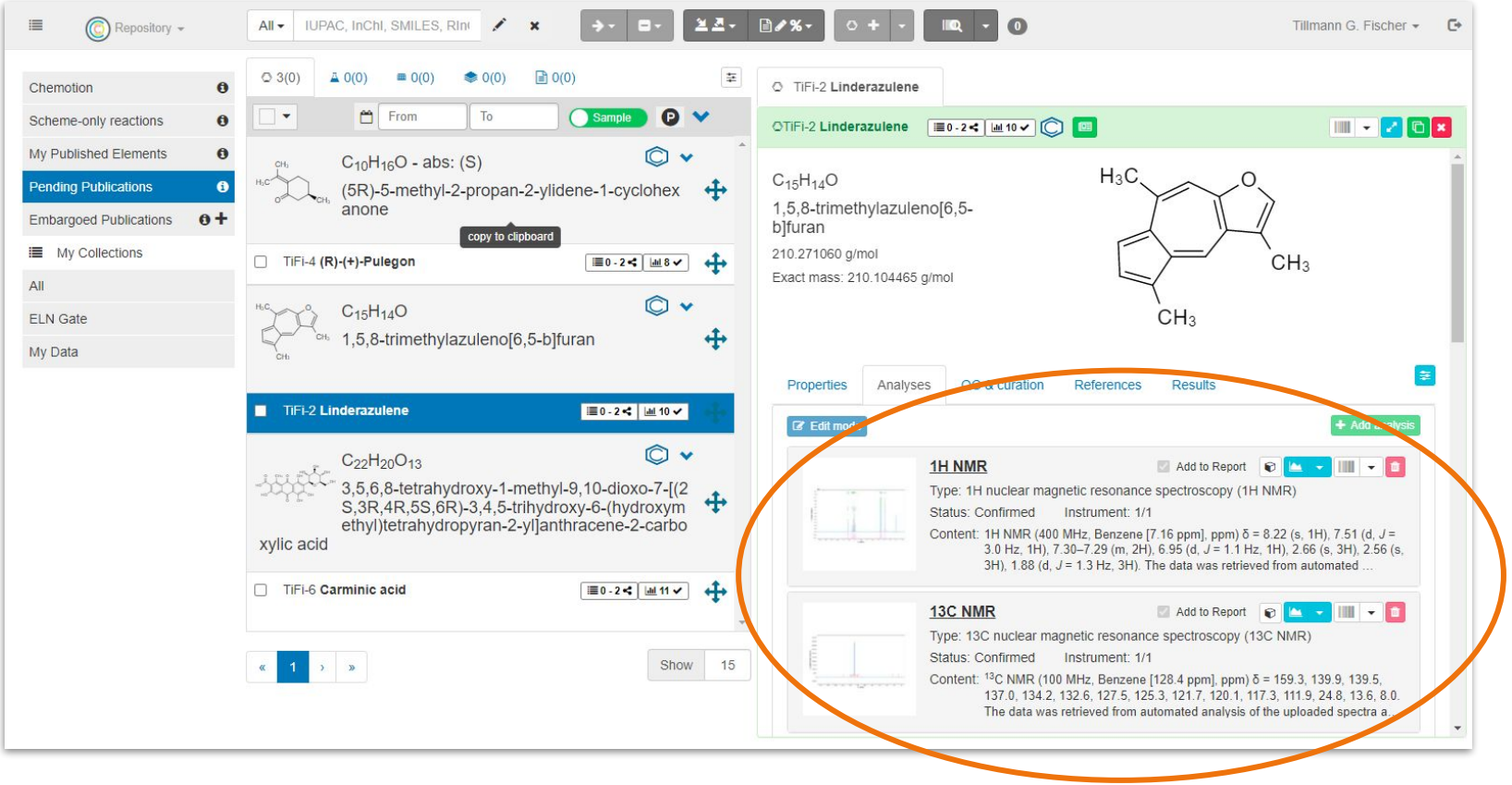
nmrium .json

jcampdx .jdx

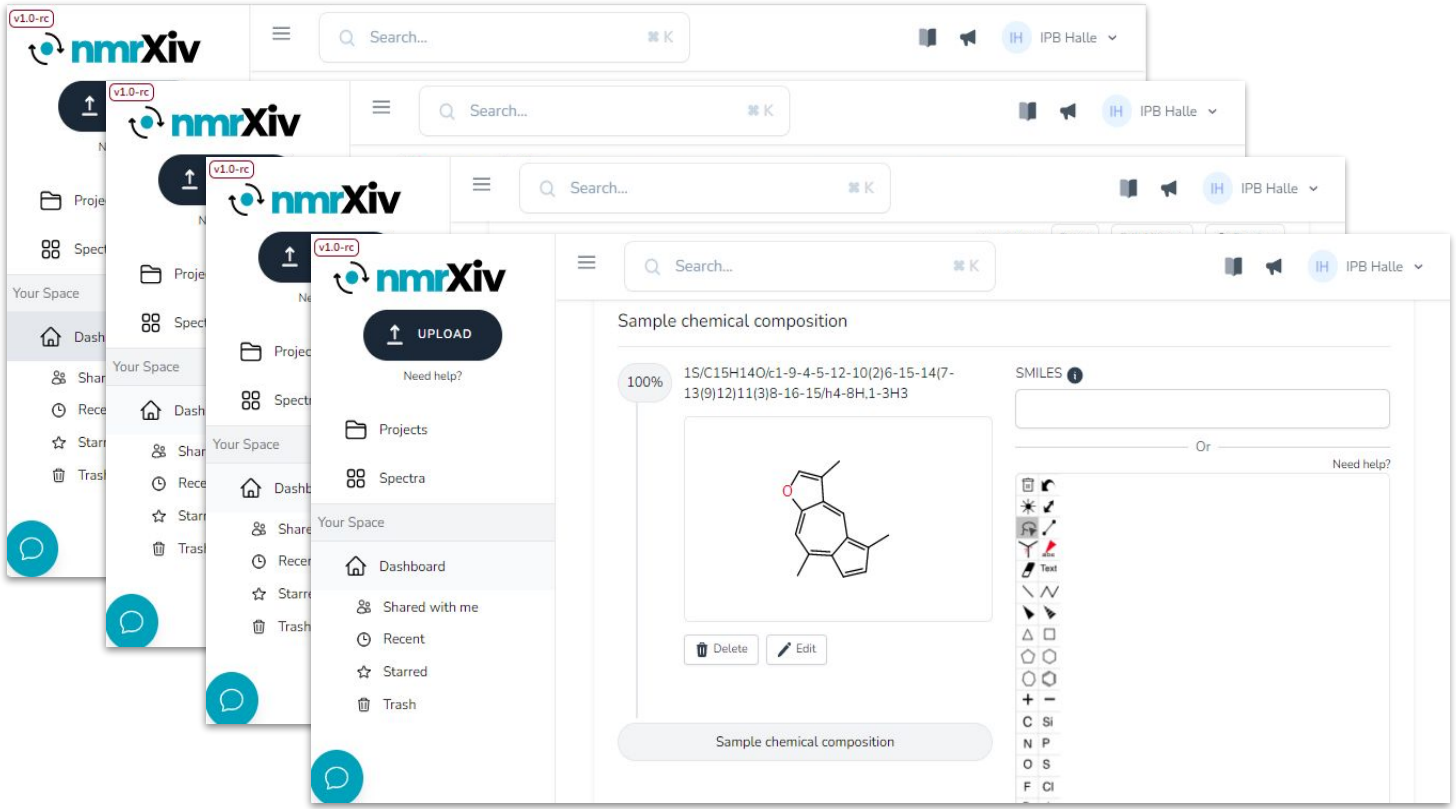
# Generische Repositorien – RADAR(4Chem)



# Feld-spezifische Repositorien – Chemotion Repository



# Feld-spezifische Repositorien – nmrXiv.org



# Beschreibende Metadaten

**Generische Metadaten**

- Dublin Core Metadaten (aus OCLC/NCSA Metadata Workshop 1995 in Dublin/Ohio):
  - …Typ, Sprache, Autor, Titel, Verleger, Rechte…
    -> Bibliotheken und Museen

- DataCite Metadaten (seit 2009. Hauptsitz an der Technischen Informationsbibliothek (TIB) Hannover):
  - DOI-Registrierungsagentur (CrossRef für wissenschaftliche Publikationen)
  - Datensätze, Software, Preprint; auch physische Objekte (PIDINST) -> siehe Dokumentation: resourceTypeGeneral

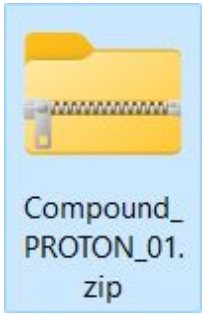
DataCite
FIND, ACCESS, AND REUSE DATA

Dublin Core Metadata, URL: https://www.dublincore.org/. Dublin Core Metadata Specification, URL: https://www.dublincore.org/specifications/dublin-core/. Dublin Core Metadata Schemas, URL: https://www.dublincore.org/schemas/. DataCite, URL: https://datacite.org/. DataCite, URL: https://datacite.org/. DataCite Metadata Schema, URL: https://schema.datacite.org/. Dublin Core under WTF Public Licence 2.0. DataCite Logo under CC BY 4.0 licence.

# (Meta-)Daten


DataCite
FIND, ACCESS, AND REUSE DATA

**Fachspezifische Metadaten**
...Struktur, Temperatur, Lösungsmittel…



# Mindestinformationstandards (MI-Standards)

"Von der wissenschaftlichen Gemeinschaft **akzeptierter Konsens über die Mindestinformationen**, die über eine chemische Untersuchung berichtet werden sollen, sowie über die Daten und Metadaten, die in einem Repository hinterlegt werden sollen."

-> **MIChI** = **M**inimum **I**nformation about a **Ch**emical **I**nvestigation

FAIR NMR Research Data Management Workshop June/2022 in Karlsruhe, **Link to Report**, Chemotion $^{1}$H NMR MI Template (GitHub, Visualisierung)

NMR MIChI? Bereits in den Messdaten enthaltene (Meta-)Daten?

# Frage an das Publikum

- *Welche NMR-Datenformaten sind bekannt?*

# Rezept für erfolgreiche Datenformate

*Zutatenliste:*

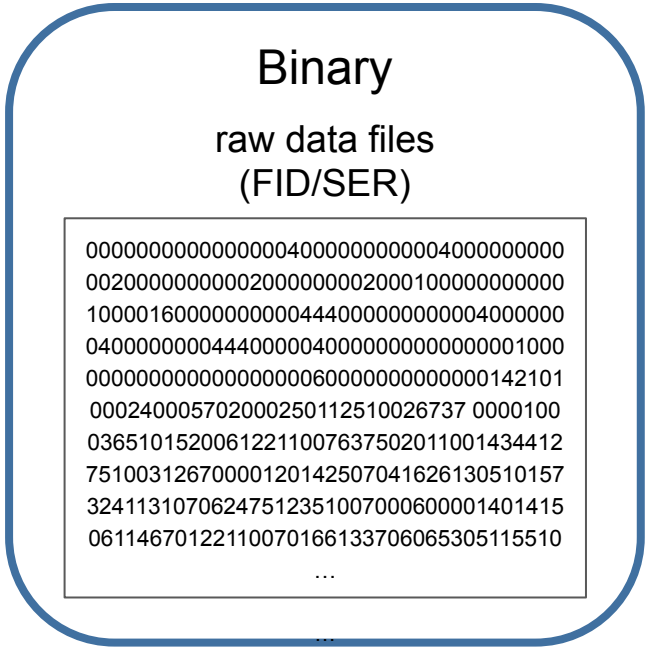
- ***Spezifikation*** *d.h. vollständige und klare Beschreibung des Formats incl. Datenmodells*
  - *Für XML und JSON auch im Format des Schemas (XSD, json.schema)*
- *Referenzimplementierung*
  - ***reader/writer, converter, validator***
  - *als Programmbibliotheke zur Implementierung in andere Software*
  - *weitere Implementierungen in anderen Programmiersprachen*
  - ***Testdaten***
- **Nutzer** und **Anbieter** von Formaten -> kann zu Henne-Ei-Problem führen

# Datenformate für NMR: Herstellerformate

Beispiel: Bruker XWIN-NMR Format

- XWIN-NMR und TopSpin
- Rohdaten (FID/SER): **Binärformat**
- Parameterdaten (acqu*/proc*): **JCAMP-DX** (Text)

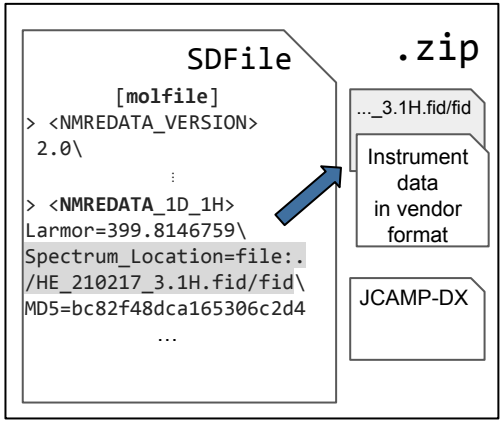
- Häufige Eigenschaften von Herstellerformaten:
  - **keine Spezifikation verfügbar**
- Referenzimplementierungen (insb. Validator) sind **keine frei nutzbare Software**



# Datenformate für NMR: Offene Formate

## NMReDATA record

Container




## nmrML

XML

```
<?xml version="1.0" encoding="UTF-8"?>
<nmrML xmlns="http://nmrml.org/schema"
xmlns:xsi="http://www.w3.org/2001/XMLSc
hema-instance" version="1.0.rc1"
xsi:schemaLocation="...nmrml.org/schema
http://nmrml.org/schema/.../nmrML.xsd">
<cvList>
<cv id="NMRCV" fullName="Nuclear
Magnetic Resonance CV"
version="1.1.0"...>
</cvList>
   <sampleList>
      <sample >
         <fieldFrequencyLock
fieldFrequencyLockName="cdcl3"/> ...
```

## JCAMP-DX

Text/Tags

```
##TITLE= testntup
##JCAMP-DX= 6.0          $$...
##DATA TYPE= NMR SPECTRUM
##DATA CLASS= NTUPLES
                    ⋮
##PAGE= N=1
##DATA TABLE=(X++(R..R)), XYDATA
16383
B254931p506547j928560N577988J225
16375
H070280o319742J809880k462623o675
16367
d539532K551632l441737j691199N043
                    ...
```

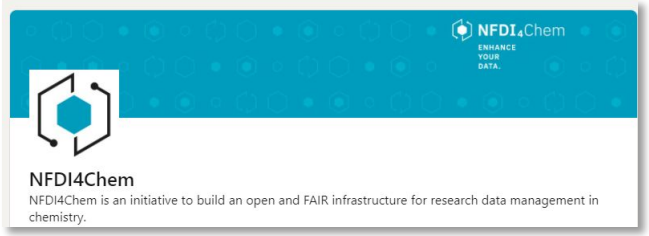

D. Rauh, C. Blankenburg, T. G. Fischer, N. Jung, S. Kuhn, U. Schatzschneider, T. Schulze, S. Neumann, *Pure Appl. Chem.* **2022**, *94*, 725. DOI: 10.1515/pac-2021-3101. M. Pupier, J.-M. Nuzillard, J. Wist, N. E. Schlörer, S. Kuhn *et al.*, NMReDATA, a standard to report the NMR assignment and parameters of organic compounds, *Magn. Reson. Chem.* **2018**, *56*, 703, DOI: 10.1002/mrc.4737. S. Kuhn, L. H. E. Wieske, P. Trevorrow, D. Schober, N. E. Schlörer *et al.*, NMReDATA: Tools and applications, *Magn. Reason. Chem.* **2018**, *59*, 792, DOI: 10.1002/mrc.5146. NMReDATA initiative, URL: http://nmredata.org/. NMReData initiative format specification, URL: http://nmredata.org/wiki/NMReDATA_tag_format_2.0. NMReData Initiative, GitHUb, URL: https://github.com/NMReDATAInitiative. Reference implementation, converter and validator in java available via GitHub. D. Schober, D. Jacob, M. Milson *et al.*, nmrML: A Community Supported Open Data Standard for the Description, Storage and Exchange of NMR Data, *Anal. Chem.* **2018**, *90*, 649. DOI: 10.1021/acs.analchem.7b02795. nmrML, https://www.nmrml.org/. nmrML/nmrML GitHub, https://github.com/nmrML/nmrML. J. Gasteiger, T. Engel, Uni Erlangen, **2004**, URL: https://www2.chemie.uni-erlangen.de/projects/vsc/chemoinformatik/erlangen/daten/jcamp.html. IUPAC, JACMP-DX, Committee on Printed and Electronic Publications (now Committee on Publications and Cheminformatics Data Standards), Subcommittee on Electronic Data Standards (now Subcommittee on Cheminformatics Data Standards), **2005**, URL: http://www.jcamp-dx.org/. A. N. Davies, R. M. Hanson, P. Lampen, R. J. Lancashire, An overview of the JCAMP-DX format, *Pure Appl. Chem.* **2022**, *94*, 705, DOI: 10.1515/pac-2021-2010. L. R. McEwen, D. Martinsen, R. Lancashire, P. Lampen, A. N. Davies, Are your spectroscopic data FAIR? *Spectrosc. Eur.* **2018**, *30*, 21, DOI: 10.1255/sew.2018.a2.

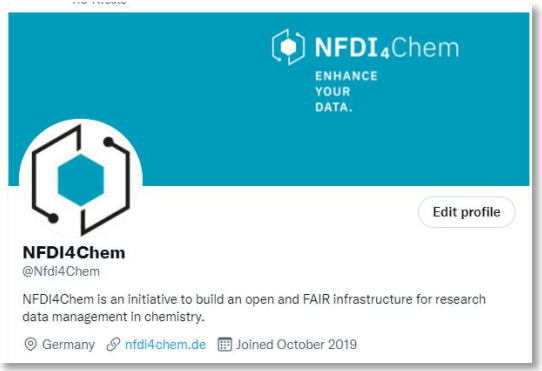
# Follow us | subscribe | stay up to date

twitter.com/**Nfdi4Chem**

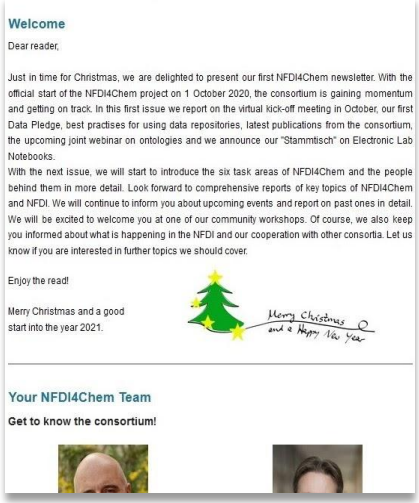
linkedin.com/**company/nfdi4chem**

Homepage: nfdi4chem.de



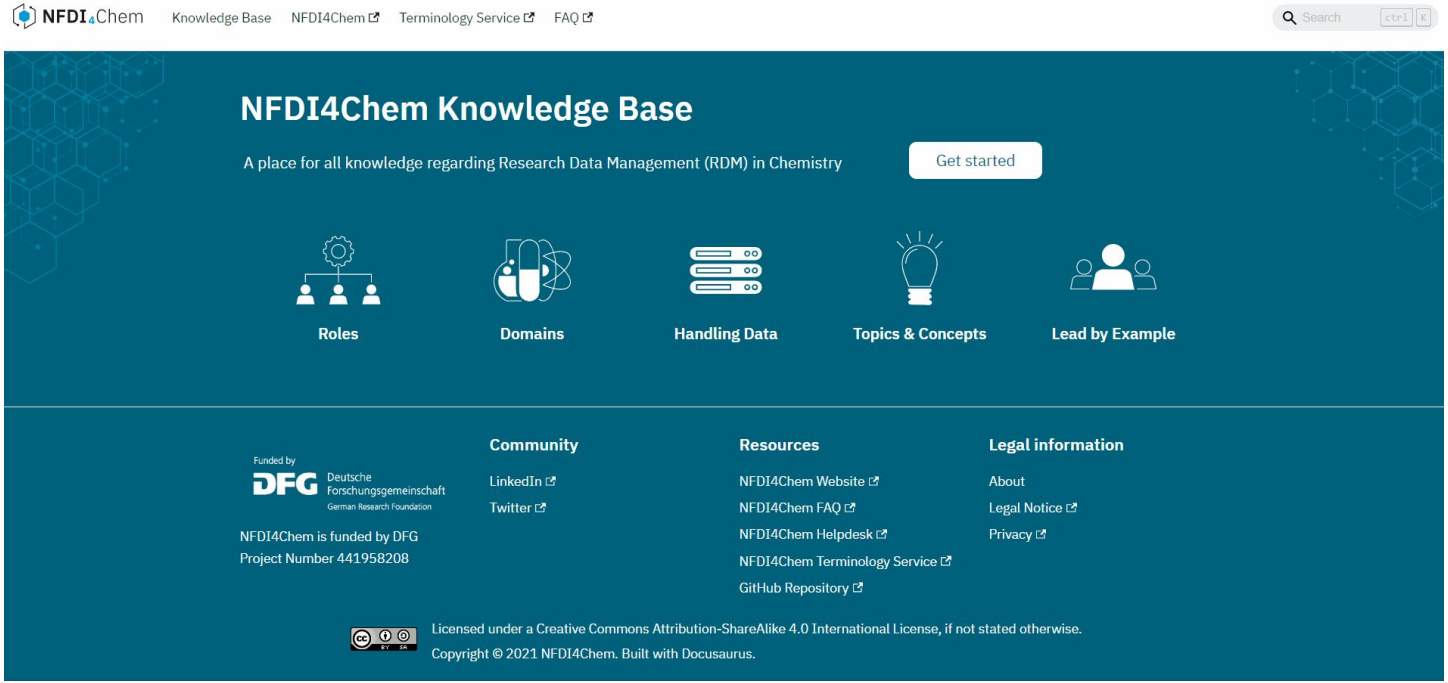





Subscribe to our newsletter which is released every 3 months via our website!

# NFDI4Chem Knowledge Base

- Access to Research Data Management (RDM) knowledge via various points of entry




NFDI4Chem Knowledge Base

# NFDI4Chem

ENHANCE
YOUR
DATA.

tfischer@ipb-halle.de

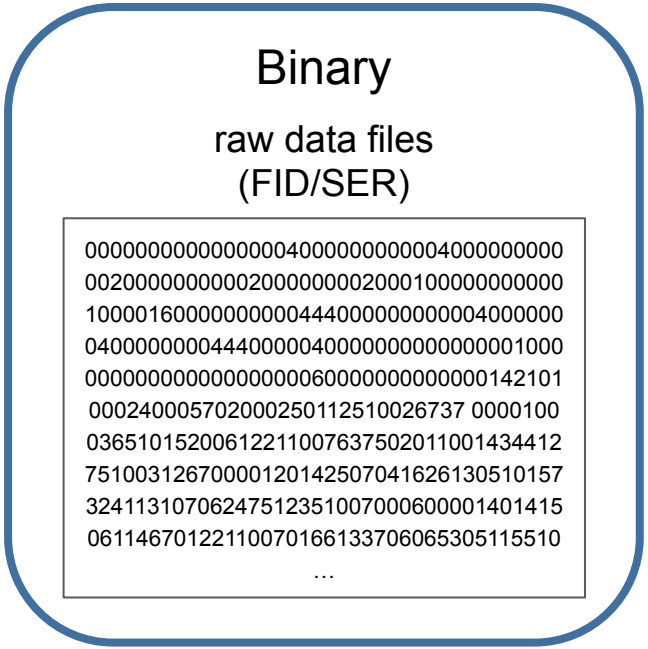
Slides: https://t1p.de/r1b6w

# Datenformate für NMR: Bruker XWIN-NMR Format

Format in XWIN-NMR und TopSpin

- Rohdaten (FID/SER): **Binärformat**
- Parameterdaten (acqu*/proc*): **JCAMP-DX** (Text)

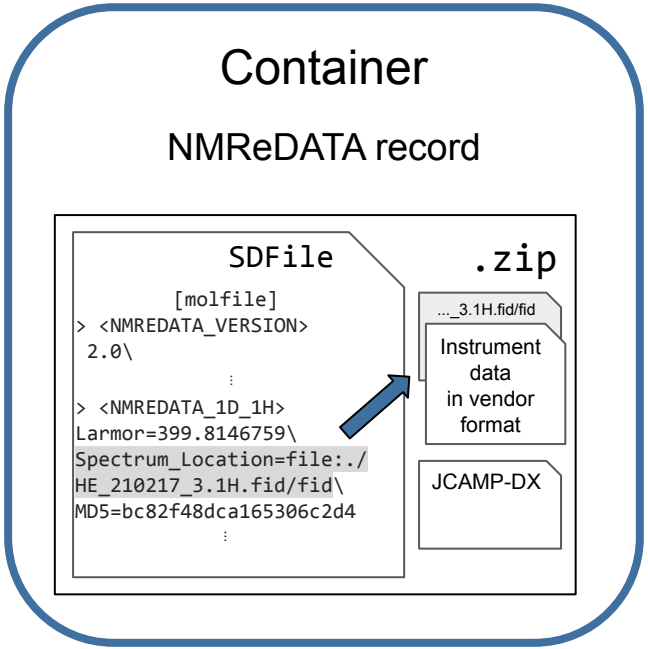
- *Data Formats User Manual* verfügbar (Beschreibung des Datenmodells), jedoch **keine Spezifikation**
- Referenzimplementierung = TopSpin, jedoch **keine frei nutzbare Software**
- **Validator nicht verfügbar**



# Datenformate für NMR: NMReDATA

NMReDATA record **Ordner/ZIP**

- Daten in **Herstellerformat**, optional zusätzlich als **JCAMP-DX**
- **SDFile** (ein CT format)
  - Struktur (molfile)
  - Chem. Verschiebungen
  - Kopplungskonstanten
  - Integrale
  - Zuordnungen,
  - Kreuzpeaks und Korrelatione
- reader/writer, converter verfügbar
- Kein Validator verfügbar




M. Pupier, J.-M. Nuzillard, J. Wist, N. E. Schlörer, S. Kuhn *et al.*, NMReDATA, a standard to report the NMR assignment and parameters of organic compounds, *Magn. Reson. Chem.* **2018**, *56*, 703, DOI: 10.1002/mrc.4737. S. Kuhn, L. H. E. Wieske, P. Trevorrow, D. Schober, N. E. Schlörer *et al.*, NMReDATA: Tools and applications, *Magn. Reason. Chem.* **2018**, *59*, 792, DOI: 10.1002/mrc.5146. NMReDATA initiative, URL: http://nmredata.org/. NMReData initiative format specification, URL: http://nmredata.org/wiki/NMReDATA_tag_format_2.0. NMReData Initiative, GitHUb, URL: https://github.com/NMReDATAInitiative. Reference implementation, converter and validator in java available via GitHub.

# Datenformate für NMR: nmrML

XML-basiertes Format

- Entwickelt für Metabolomics Daten
  - Format für NMR daten im Repositorium Metabolights
- **Spezifikation, XSD Datei, XSD Dokumentation und nmrCV Ontologie verfügbar**
- **reader/writer, converter und Validator verfügbar**

XML

**e**xtensible **m**arkup **l**anguage

nmrML

```
<?xml version="1.0" encoding="UTF-8"?>
<nmrML xmlns="http://nmrml.org/schema"
xmlns:xsi="http://www.w3.org/2001/XMLSche
ma-instance" version="1.0.rc1"
xsi:schemaLocation="...nmrml.org/schema
http://nmrml.org/schema/.../nmrML.xsd">
<cvList>
<cv id="NMRCV" fullName="Nuclear Magnetic
Resonance CV" version="1.1.0"...>
</cvList>
   <sampleList>
      <sample >
         <fieldFrequencyLock
fieldFrequencyLockName="cdcl3"/>
```

D. Rauh, C. Blankenburg, T. G. Fischer, N. Jung, S. Kuhn, U. Schatzschneider, T. Schulze, S. Neumann, *Pure Appl. Chem.* **2022**, *94*, 725. DOI: 10.1515/pac-2021-3101. D. Schober, D. Jacob, M. Milson *et al.*, nmrML: A Community Supported Open Data Standard for the Description, Storage and Exchange of NMR Data, *Anal. Chem.* **2018**, *90*, 649. DOI: 10.1021/acs.analchem.7b02795. nmrML, https://www.nmrml.org/. nmrML/nmrML GitHub, https://github.com/nmrML/nmrML.

# Datenformate für NMR: JCAMP-DX

```
##TITLE= ACET.SPOROPO.TYPHA/CH-CORR
##JCAMP-DX= 6.00   $$ ISAS JCAMP-DX program (Version 2.0 beta)
##DATA TYPE= NMR SPECTRUM
##DATA CLASS= NTUPLES
##NUM DIM= 2
##ORIGIN= Dr. J. Lambert, ISAS Dortmund, Germany
##OWNER= COPYRIGHT (C) 1999 by ISAS Dortmund, Germany
##.SOLVENT NAME= CDCL3
##SPECTROMETER/DATA SYSTEM= JEOL GX 400
##.FIELD= 9.1
##.OBSERVE FREQUENCY=  100.40
##.OBSERVE NUCLEUS= ^13C
##.PULSE SEQUENCE= HETCOR
s1=         0.0
s3=         0.0
s4=        14
p1=        10.500
p2=        21.000
p3=        28.700
rd=      1000.000
id1=        0.114
d1=         0.114
d2=        10.000
NS=       2720
##.ACQUISITION SCHEME= NOT PHASE SENSITIVE
##.DIGITISER RES= 16
##.ZERO FILL=       0
##RESOLUTION=     5.87
```

Text/Tags with encoded data

JCAMP-DX

```
##TITLE= testntup
##JCAMP-DX= 6.0          $$ Bruker...
##DATA TYPE= NMR SPECTRUM
##DATA CLASS= NTUPLES
                    ⋮
##PAGE= N=1
##DATA TABLE=(X++(R..R)), XYDATA
16383 B254931p506547j928560N577988J225
16375 H070280o319742J809880k462623o675
16367 d539532K5516321441737j691199N043
                    ⋮
```

*Textbasiertes Format. Gekennzeichnete Dateneinträge (labelled data records = LDR): ##Deskriptor= Wert*

# Datenformate für NMR: JCAMP-DX

**Table 2:** Example of an IR data file.

```
##TITLE= o-eugenol
##JCAMP-DX= 5.01 $$export from JSpecView
##DATA TYPE= INFRARED SPECTRUM
##DATA CLASS= XYDATA
##ORIGIN= Dept of Chem, UWI, JAMAICA
##OWNER= public domain
##LONGDATE= 1997/07/01 09:24:00
##SPECTROMETER DATA SYSTEM= PERKIN-ELMER 1000 FT-IR
##INSTRUMENTAL PARAMETERS= 4400.00,450.00 cm-1; 16 scans; mode ratio; apod strong
##STATE= liquid
##IUPAC NAME= 2-allyl-6-methoxyphenol
##MOLFORM= C10 H12 O2
##BP= 271
##RESOLUTION= 2
##XUNITS= 1/CM
##YUNITS= TRANSMITTANCE
##XFACTOR= 3.93E-03
##YFACTOR= 9.40262357E-07
##FIRSTX= 470
##FIRSTY= 0.00520004
##LASTX= 4400
##NPOINTS= 3931
##XYDATA= (X++(Y...Y)) 119593E530R12K042n65l78n02q80J885K545o2j634K702r4r43J760J759O29K859A7942
124427B0990L048J89k891K263L362J288K325L237N97l14K891K828O91K043L865D8454
128753E1031K733M305J666K83L802M902J319L048N311J603n35K703K891N122M022I5934
133079I8416M431L991M336L174L362M493K766N813N844N374N122O630Q798O945A74114
137150A85238J0401q673m682m148m745l519l268k514k828l017l173m368m714o096A32353
.....
1119593H75220 $$checkpoint
##END=
```

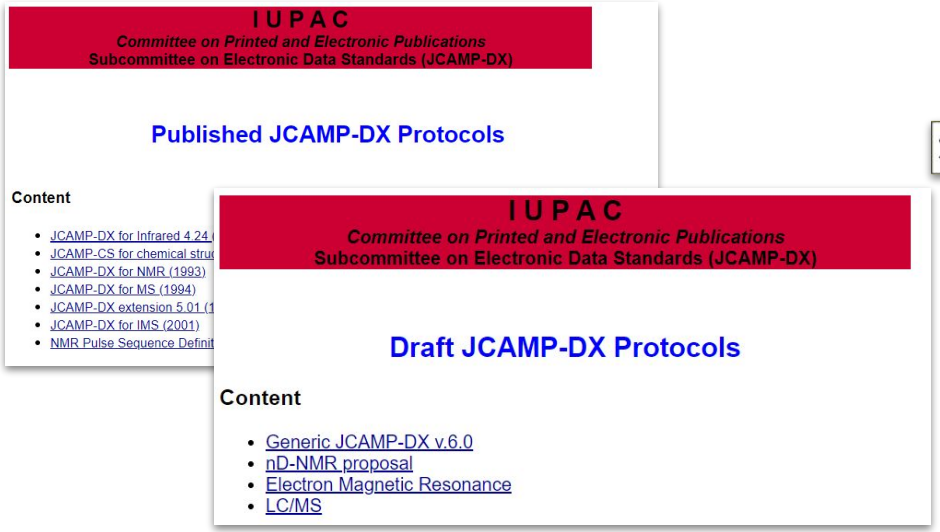
**Table 6:** Example Bruker NMR JCAMP-DX file showing private labels and comments.

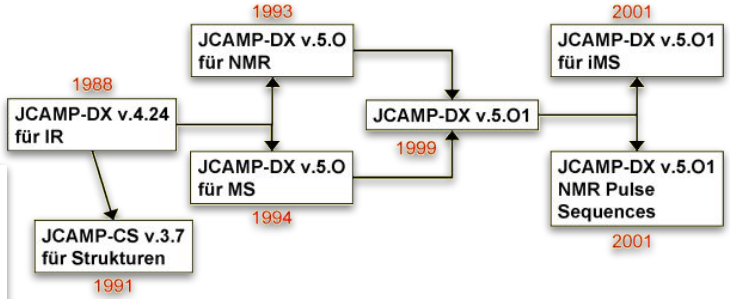
```
$$ Bruker specific parameters
$$ ————————————————————————
##$DATPATH= <C:\Bruker\data\cga\nmr>
##$EXPNO= 28
##$NAME= <cyclosporine-0813>
##$PROCNO= 1
##$ACQT0= −6.36618283677107
##$AMP= (0...31)
100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
100 100 100 100 100 100 100 100 100 100 100 100 100 100 100 100
##$AMPCOIL= (0...19)
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
##$ANAVPT= −1
##$AQSEQ= 0
##$AQ_mod= 3
##$AUNM= <au_sel180zg>
##$AUTOPOS= <>
.....
$$ End of Bruker specific parameters
$$ ————————————————————————————
```

*Ausgewählte Erweiterungen:*

- *Private Dateneinträge (private data labels) ##$Descriptor = Wert*
- *Kommentare (comments) $$ (one line)*

# Datenformate: JCAMP-DX



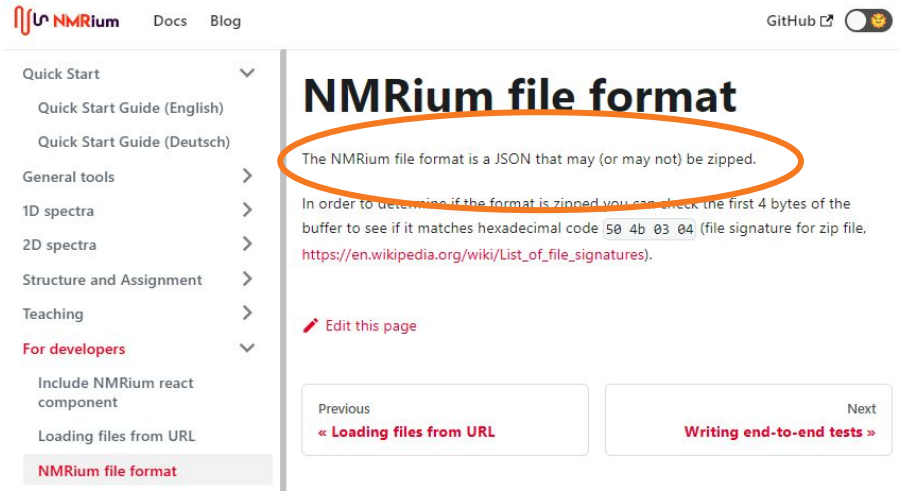




*Weiterentwicklung von JCAMP-DX 2001 mit IUPAC gestoppt zur Vereinigung der Kräfte zur Entwicklung eines XML basierten Formats (AniML).*

*Entwurf für 2D NMR Daten -> implementiert in TopSpin.*


J. Gasteiger, T. Engel, Uni Erlangen, **2004**, URL: https://www2.chemie.uni-erlangen.de/projects/vsc/chemoinformatik/erlangen/daten/jcamp.html. D. Rauh, C. Blankenburg, T. G. Fischer, N. Jung, S. Kuhn, U. Schatzschneider, T. Schulze, S. Neumann, Data format standards in analytical chemistry, *Pure Appl. Chem.* **2022**, *94*, 725, DOI: 10.1515/pac-2021-3101. IUPAC, JACMP-DX, Committee on Printed and Electronic Publications (now Committee on Publications and Cheminformatics Data Standards), Subcommittee on Electronic Data Standards (now Subcommittee on Cheminformatics Data Standards), **2005**, URL: http://www.jcamp-dx.org/. A. N. Davies, R. M. Hanson, P. Lampen, R. J. Lancashire, An overview of the JCAMP-DX format, *Pure Appl. Chem.* **2022**, *94*, 705, DOI: 10.1515/pac-2021-2010. L. R. McEwen, D. Martinsen, R. Lancashire, P. Lampen, A. N. Davies, Are your spectroscopic data FAIR? *Spectrosc. Eur.* **2018**, *30*, 21, DOI: 10.1255/sew.2018.a2.

# Datenformate für NMR: NMRium format




- in JSON

JSON

JavaScript Object Notation

```
{
    "$schema": "http://.../”
    "title" : "..."
    "description" : "JSON-..."
    "type" : "object",
    "properties" : {
        "@id": { "type": "string",
        "filename": { "type" : "st
        "identifier" : { "type" :
        "title" : { "type" : "stri
        "description" : { "type" :
                    ⋮
}
```

NMRium Documentation, NMRium file format, URL: https://docs.nmrium.org/for-developers/nmrium_file_format/.