The Chemical Analysis Use Case

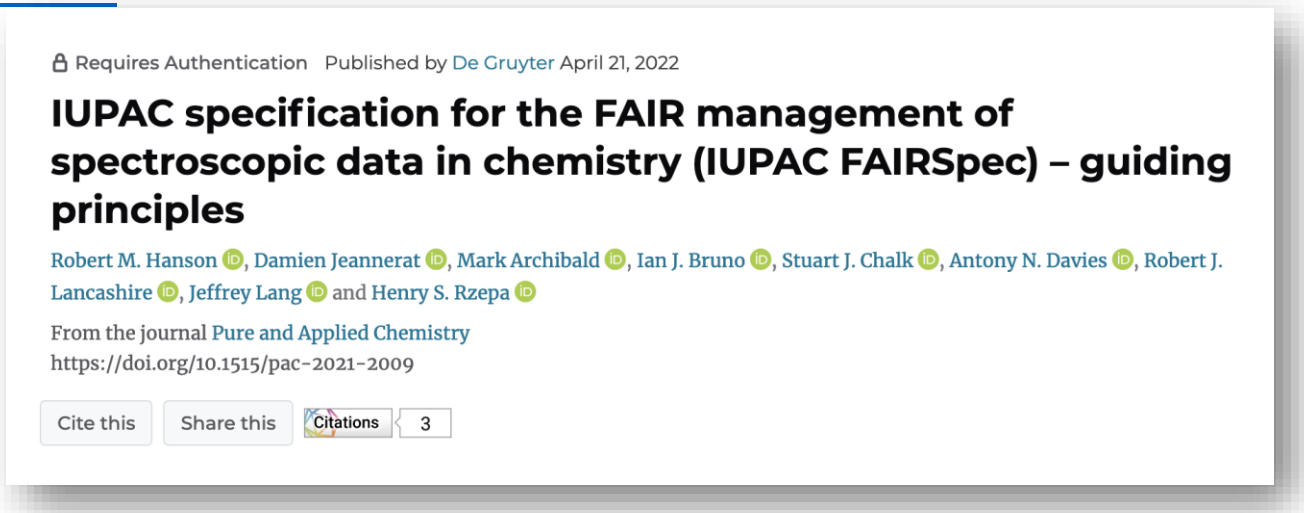
Stuart Chalk, Department of Chemistry, University of North Florida
IUPAC CPCDS, IUPAC FAIRSpec Project, WorldFAIR WP3, CODATA DRUM TG

- General thoughts: Chemical analysis in the laboratory
  - Plethora of instrument types, complexities, scale of data generation
  - Instrument components: how many need unique PIDs? -> peripherals
  - Contextualizing a measurement
    - Environmental conditions: temperature, humidity, power stability
    - Calibration: Most recent, history, trends in results -> contamination
    - Service: Date of last service
  - Problems with instrument data
    - No units! Data not reported with the correct accuracy (significant digits)
    - Proprietary software hides useful data in some cases
    - Data export may not be complete to open formats (e.g., JCAMP-DX)

https://iupac.org/what-we-do/digital-standards/jcamp-dx/

IUPAC FAIRSpec Project

- The project will develop standards for the production and dissemination of digital data objects that contain enough spectral data and metadata that they can be (a) findable through semantic searches on the web, (b) available through standard interfaces, (c) interoperable and transferable between systems, and (d) readable and reusable over time, for both humans and machines.
- Working on a finding aid and package format to link chemical substances to spectral data
- GitHub Repository: https://github.com/IUPAC/IUPAC-FAIRSpec
- Project Website: https://iupac.org/project/2019-031-1-024/
- Paper: https://doi.org/10.1515/pac-2021-2009



Requires Authentication Published by De Gruyter April 21, 2022

IUPAC specification for the FAIR management of spectroscopic data in chemistry (IUPAC FAIRSpec) – guiding principles

Robert M. Hanson, Damien Jeannerat, Mark Archibald, Ian J. Bruno, Stuart J. Chalk, Antony N. Davies, Robert J. Lancashire, Jeffrey Lang and Henry S. Rzepa

From the journal Pure and Applied Chemistry
https://doi.org/10.1515/pac-2021-2009

Units of Measurement Interoperability

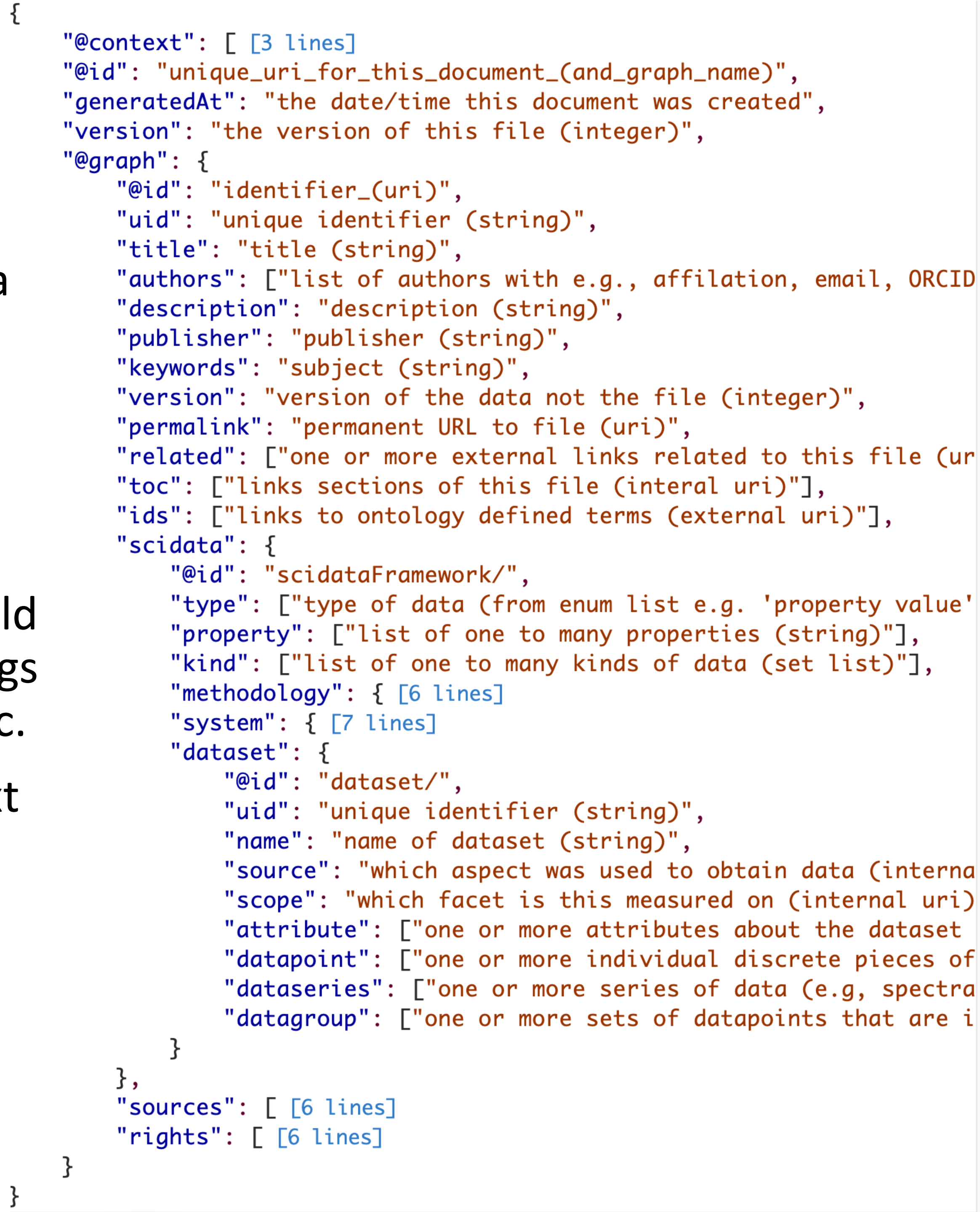
- In Nov. 2022 the 27th CGPM (Committee on the Metre Convention) approved a resolution to digitalize the SI (work to be done 2024-2027)
- Two important services to support this effort are already under development
  - An SI Digital Reference: A digital version of the content of the SI (with unique endpoints for all the units and prefixes currently defined in the SI brochure)
  - The Units of Measurement Interoperability Service (UMIS): A mapping service between different digital unit representation systems (there are ~20) and reference back to the SI Digital Reference.
- CIPM TG on the Digital SI: https://www.bipm.org/en/committees/ci/cipm/wg/cipm-tg-dsi
- CODATA TG Digital Representation of Units of Measurement (DRUM): https://codata.org/initiatives/task-groups/drum/
- Inventory of units of measurement representation systems: https://codata.org/wp-content/uploads/2022/12/DRUM_Units_Inventory_120522.pdf

COMMITTEE ON DATA
CODATA
INTERNATIONAL
SCIENCE COUNCIL

An Open Semantic Framework for Data-Driven Discovery

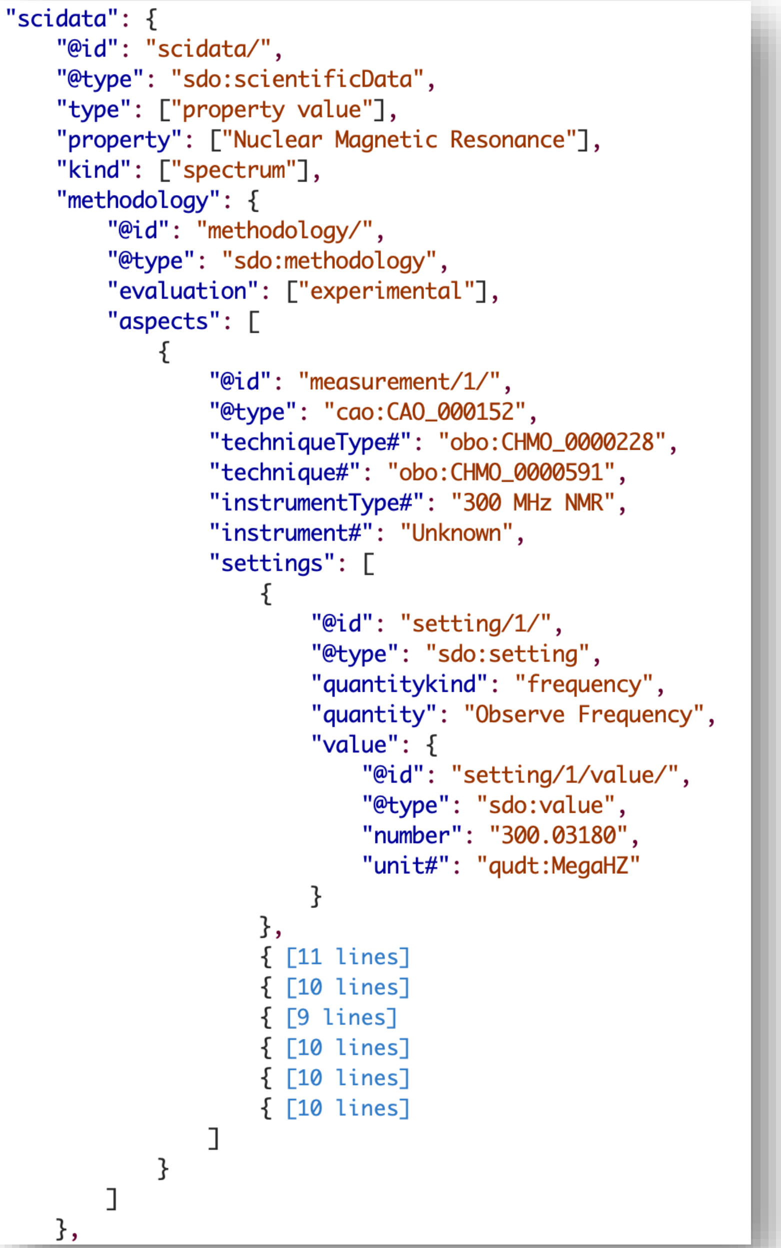
- SciData: A generic framework for storing research data
- Based on three categories of data:
  Methodology (how), System (what) and Dataset
- JSON-LD as the encoding supported by the SciData Ontology that defines the structure of data in RDF
- Methodology and System are generic containers to hold contextual data about instruments, procedures, settings and what has been studied – chemicals, organisms, etc.
- The more detailed the metadata the better the context of the research is represented
- More info at: https://stuchalk.github.io/scidata/

Award: 1835643

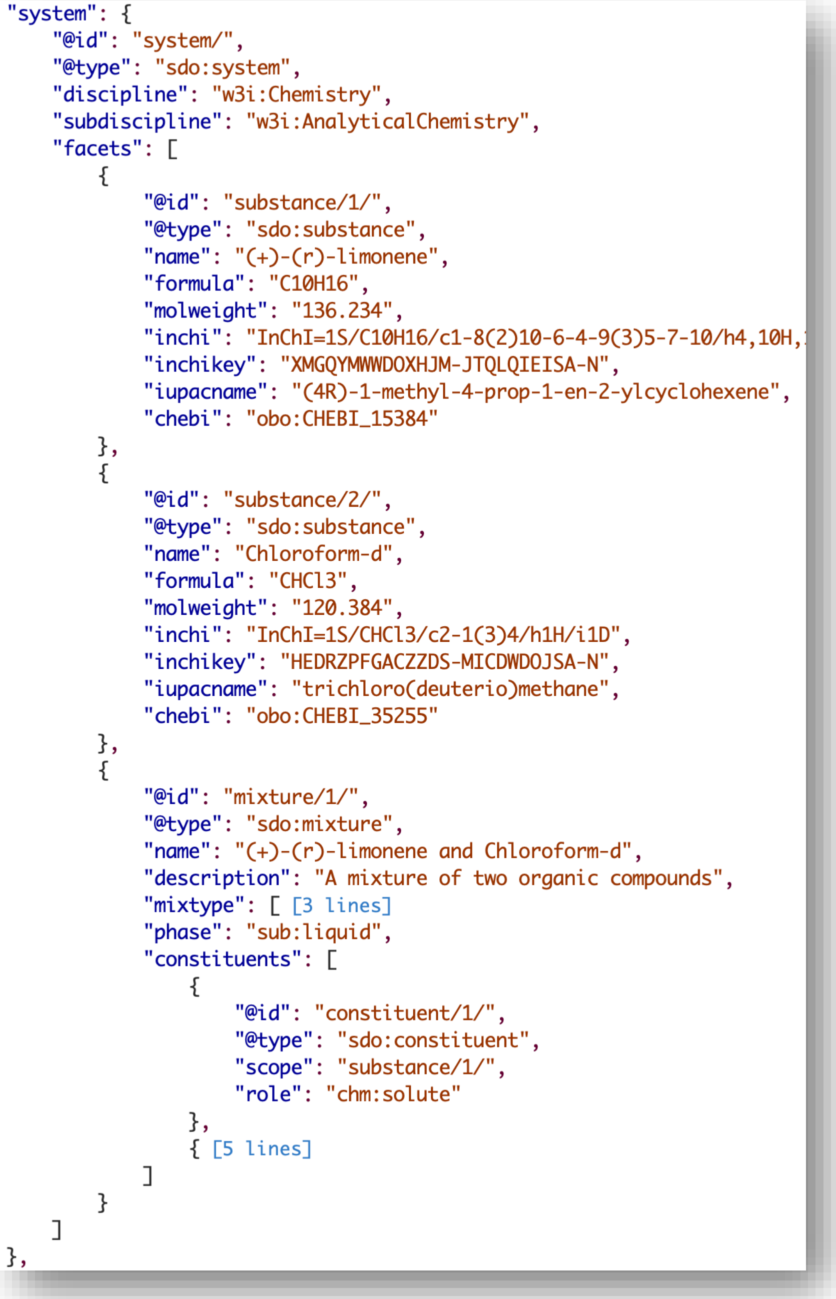


```
{
  "@context": [ [3 lines]
  "@id": "unique_uri_for_this_document_(and_graph_name)",
  "generatedAt": "the date/time this document was created",
  "version": "the version of this file (integer)",
  "@graph": {
    "@id": "identifier_(uri)",
    "uid": "unique identifier (string)",
    "title": "title (string)",
    "authors": ["list of authors with e.g., affilation, email, ORCID
    "description": "description (string)",
    "publisher": "publisher (string)",
    "keywords": "subject (string)",
    "version": "version of the data not the file (integer)",
    "permalink": "permanent URL to file (uri)",
    "related": ["one or more external links related to this file (ur
    "toc": ["links sections of this file (interal uri)"],
    "ids": ["links to ontology defined terms (external uri)"],
    "scidata": {
      "@id": "scidataFramework/",
      "type": ["type of data (from enum list e.g. 'property value'
      "property": ["list of one to many properties (string)"],
      "kind": ["list of one to many kinds of data (set list)"],
      "methodology": { [6 lines]
      "system": { [7 lines]
      "dataset": {
        "@id": "dataset/",
        "uid": "unique identifier (string)",
        "name": "name of dataset (string)",
        "source": "which aspect was used to obtain data (interna
        "scope": "which facet is this measured on (internal uri)
        "attribute": ["one or more attributes about the dataset
        "datapoint": ["one or more individual discrete pieces of
        "dataseries": ["one or more series of data (e.g, spectra
        "datagroup": ["one or more sets of datapoints that are i
      }
    },
    "sources": [ [6 lines]
    "rights": [ [6 lines]
  }
}
```

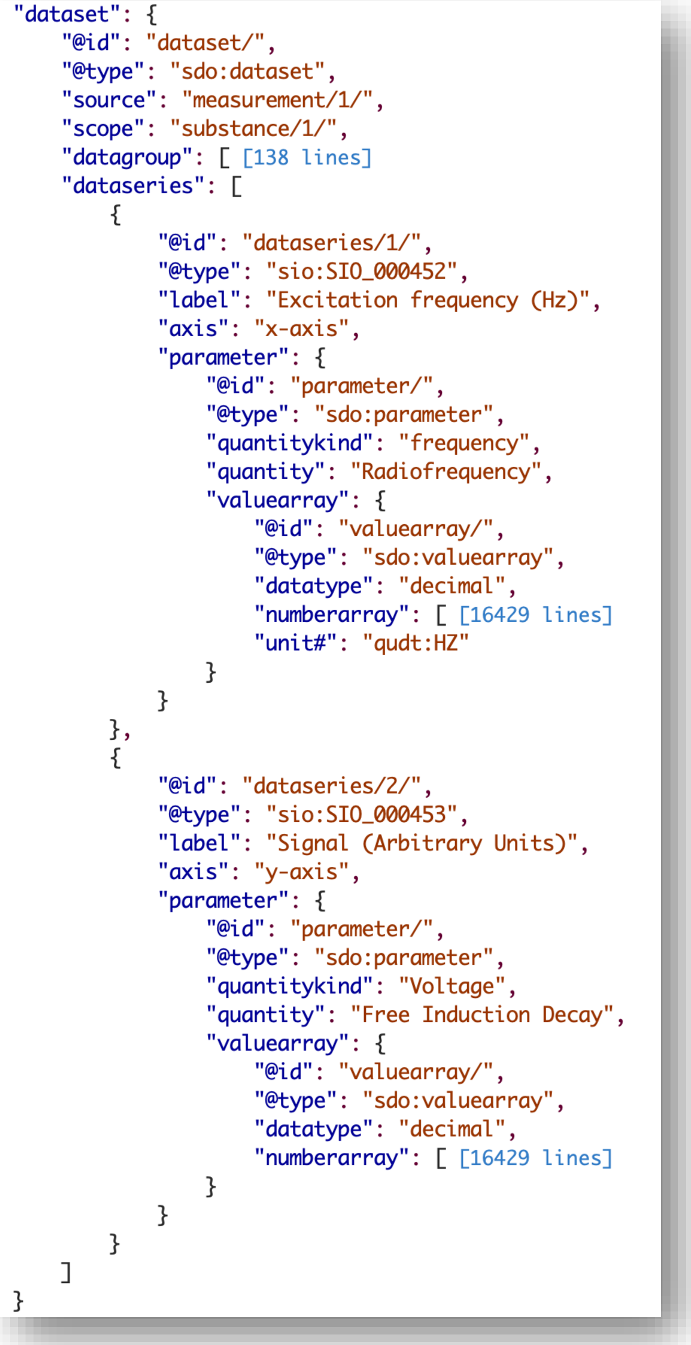
SciData Example



```
"scidata": {
    "@id": "scidata/",
    "@type": "sdo:scientificData",
    "type": ["property value"],
    "property": ["Nuclear Magnetic Resonance"],
    "kind": ["spectrum"],
    "methodology": {
        "@id": "methodology/",
        "@type": "sdo:methodology",
        "evaluation": ["experimental"],
        "aspects": [
            {
                "@id": "measurement/1/",
                "@type": "cao:CAO_000152",
                "techniqueType#": "obo:CHMO_0000228",
                "technique#": "obo:CHMO_0000591",
                "instrumentType#": "300 MHz NMR",
                "instrument#": "Unknown",
                "settings": [
                    {
                        "@id": "setting/1/",
                        "@type": "sdo:setting",
                        "quantitykind": "frequency",
                        "quantity": "Observe Frequency",
                        "value": {
                            "@id": "setting/1/value/",
                            "@type": "sdo:value",
                            "number": "300.03180",
                            "unit#": "qudt:MegaHZ"
                        }
                    },
                    { [11 lines]
                    { [10 lines]
                    { [9 lines]
                    { [10 lines]
                    { [10 lines]
                    { [10 lines]
                ]
            }
        ]
    },
```



```
"system": {
    "@id": "system/",
    "@type": "sdo:system",
    "discipline": "w3i:Chemistry",
    "subdiscipline": "w3i:AnalyticalChemistry",
    "facets": [
        {
            "@id": "substance/1/",
            "@type": "sdo:substance",
            "name": "(+)-(r)-limonene",
            "formula": "C10H16",
            "molweight": "136.234",
            "inchi": "InChI=1S/C10H16/c1-8(2)10-6-4-9(3)5-7-10/h4,10H,
            "inchikey": "XMGQYMWWDOXHJM-JTQLQIEISA-N",
            "iupacname": "(4R)-1-methyl-4-prop-1-en-2-ylcyclohexene",
            "chebi": "obo:CHEBI_15384"
        },
        {
            "@id": "substance/2/",
            "@type": "sdo:substance",
            "name": "Chloroform-d",
            "formula": "CHCl3",
            "molweight": "120.384",
            "inchi": "InChI=1S/CHCl3/c2-1(3)4/h1H/i1D",
            "inchikey": "HEDRZPFGACZZDS-MICDWDOJSA-N",
            "iupacname": "trichloro(deuterio)methane",
            "chebi": "obo:CHEBI_35255"
        },
        {
            "@id": "mixture/1/",
            "@type": "sdo:mixture",
            "name": "(+)-(r)-limonene and Chloroform-d",
            "description": "A mixture of two organic compounds",
            "mixtype": [ [3 lines]
            "phase": "sub:liquid",
            "constituents": [
                {
                    "@id": "constituent/1/",
                    "@type": "sdo:constituent",
                    "scope": "substance/1/",
                    "role": "chm:solute"
                },
                { [5 lines]
            ]
        }
    ]
},
```



```
"dataset": {
    "@id": "dataset/",
    "@type": "sdo:dataset",
    "source": "measurement/1/",
    "scope": "substance/1/",
    "datagroup": [ [138 lines]
    "dataseries": [
        {
            "@id": "dataseries/1/",
            "@type": "sio:SIO_000452",
            "label": "Excitation frequency (Hz)",
            "axis": "x-axis",
            "parameter": {
                "@id": "parameter/",
                "@type": "sdo:parameter",
                "quantitykind": "frequency",
                "quantity": "Radiofrequency",
                "valuearray": {
                    "@id": "valuearray/",
                    "@type": "sdo:valuearray",
                    "datatype": "decimal",
                    "numberarray": [ [16429 lines]
                    "unit#": "qudt:HZ"
                }
            }
        },
        {
            "@id": "dataseries/2/",
            "@type": "sio:SIO_000453",
            "label": "Signal (Arbitrary Units)",
            "axis": "y-axis",
            "parameter": {
                "@id": "parameter/",
                "@type": "sdo:parameter",
                "quantitykind": "Voltage",
                "quantity": "Free Induction Decay",
                "valuearray": {
                    "@id": "valuearray/",
                    "@type": "sdo:valuearray",
                    "datatype": "decimal",
                    "numberarray": [ [16429 lines]
                }
            }
        }
    ]
}
```

https://stuchalk.github.io/scidata/examples/nmr.jsonld