Aberystwyth University

Proposed Reporting Requirements for the Description of NMR-Based Metabolomics Experiments

Rubtsov, Denis V.; Jenkins, Helen; Ludwig, Christian; Easton, John; Viant, Mark R.; Günther, Ulrich; Griffin, Julian L.; Hardy, Nigel

Published in:
Metabolomics

DOI:
10.1007/s11306-006-0040-4

Publication date:
2007

Citation for published version (APA):
Rubtsov, D. V., Jenkins, H., Ludwig, C., Easton, J., Viant, M. R., Günther, U., Griffin, J. L., & Hardy, N. (2007). Proposed Reporting Requirements for the Description of NMR-Based Metabolomics Experiments. *Metabolomics*, *3*(3), 223-229. https://doi.org/10.1007/s11306-006-0040-4

PROPOSED REPORTING REQUIREMENTS FOR THE DESCRIPTION OF NMR-BASED METABOLOMICS EXPERIMENTS

Denis V. Rubtsov*[1], Helen Jenkins*[2], Christian Ludwig[3], John Easton[4], Mark R. Viant[5], Ulrich Günther[3], Julian L. Griffin[1] and Nigel Hardy[2]

[1]*Department of Biochemistry, University of Cambridge, The Hopkins Building, Cambridge CB2 1QW, UK*

[2]*Department of Computer Science, University of Wales, Aberystwyth, Ceredigion, SY23 3DB, UK*

[3]*The Henry Wellcome Building for Biomolecular NMR Spectroscopy, University of Birmingham, Birmingham, B15 2TT, UK*

[4]*School of Engineering, University of Birmingham, Birmingham, B15 2TT, UK*

[5]*School of Biosciences, University of Birmingham, Birmingham, B15 2TT, UK*

* These authors have contributed equally to the project.

* Corresponding authors:

D.V. Rubtsov: Department of Biochemistry, University of Cambridge, The Hopkins Building, Building O, Downing Site, Cambridge CB2 1QW, UK

Tel. +44 (0)1223 764948 Fax 01223 333345, Email: dvr22@cam.ac.uk.

H. Jenkins: Department of Computer Science, University of Wales, Aberystwyth, Ceredigion, SY23 3DB, Wales, UK, Tel: +44 (0)1970 622438, Email: haf@aber.ac.uk

Electronic Supplementary Material Supplementary material is available for this article at http://dx.doi.org/DOI and is accessible for authorized users

Abbreviated title: Reporting requirements for NMR-based metabolomics experiments

Acknowledgements

This project was partially funded by BBSRC grant BB/B500107/1 and the BBSRC funded UK Centre for Plant and Microbial Metabolomics (grant MET20483). Mark R. Viant thanks the Natural Environment Research Council, UK, for an Advanced Fellowship in Metabolomics (NER/J/S/2002/00618).

Abstract

The amount of data generated by NMR-based metabolomic experiments is increasing rapidly. Furthermore, diverse techniques increase the need for informative and comprehensive meta-data. These factors present a challenge in the dissemination, interpretation, reviewing and comparison of experimental results using this technology. Thus, there is a strong case for unification and standardization of the data representation for both academia and industry. Here, a systems analysis of an NMR-based metabolomics experiment is presented in order to reveal the reporting requirements. An in-depth analysis of the NMR component of a metabolomics experiment has been produced, and a first round of data standard development completed. This has focussed on both one- and two-dimensional ^{1}H NMR experiments, but is also applicable to higher dimensions and other nuclei. We also report the modelling of this schema using Unified Modelling Language (UML), and have extended this to a proof-of-concept implementation of the standard as an XML schema.

Keywords: metabolomics, NMR, standardization, reporting requirements, data model

1. Introduction

The use of Nuclear Magnetic Resonance (NMR) spectroscopy based metabolomics has increased dramatically in the last few years in a range of fields including functional genomics, toxicology, and environmental and nutritional studies (Nicholson et al., 2002; Lindon et al., 2004; Lindon et al., 2003; Griffin et al., 2001; Nicholson et al., 2005; Griffin et al., 2004; Viant et al., 2003). One- and two-dimensional (1D and 2D) ^{1}H NMR of solution state biofluids or tissue extracts have become some of the most popular tools used for metabolomics, benefiting from being high throughput in nature, relatively cheap on a per sample basis and potentially non invasive. With improvements in automation and flow probe technology, sample throughput for metabolite rich fluids such as urine and plasma is set to increase. Using such approaches, biofluid analysis has recently been used to generate predictive pattern recognition models for detecting early stage hepato- and renal toxicity following the acquisition of data on 150 model hepatic and renal toxins (Lindon et al., 2003).

As NMR spectroscopy is a high-throughput technique, the amount of data generated by this approach is increasing rapidly. The context-dependent nature of information in metabolomics studies adds complexity to the problem of systematically describing the experiment results. That is, meaningful interpretation of the results of metabolomics experiments is possible only in a specific experimental context that needs to be captured. The presence of many potential, and not always obvious, sources of experimental variation makes it difficult to extract the relevant biological information contained within metabolomic data and requires a detailed experiment description. This presents a challenge to the dissemination, interpretation, reviewing and comparison of these experimental results. Moreover, since several different NMR experiments are used in metabolomics, comprehensive and metadata-rich description plays a crucial role in facilitating adequate cross-comparison and assessment of results.

Thus there is a strong case for unification and standardization of data representation for NMR-based metabolomics. Indeed, this problem is not limited to metabolomics in general, or NMR-based metabolomics in particular, but also affects the reporting of other functional genomics datasets. In answer to this demand several initiatives have emerged, including the MGED (Microarray Gene Expression Data Society) for transcriptomics (http://www.mged.org)(Spellman et al., 2002), the Proteomics Standards Initiative (http://psidev.sourceforge.net/gps/index.html) for proteomics and the FuGE (Functional Genomics Experiment) project (http://fuge.sourceforge.net/index.php) for functional genomics. For metabolomics, SMRS (Standardisation of Reporting Methods for Metabolic Analysis) (Lindon et al., 2005), MIAMET (Minimum Information about a Metabolomics Experiment)(Bino et al., 2004) and ArMet (Architecture for Metabolomics) (Jenkins et al., 2004) were all developed in parallel and are now serving as inputs to the Metabolomics Standardisation Initiative (MSI) that is being orchestrated by the Metabolomics Society (http://www.metabolomicssociety.org/mstandards.html). However, these latter initiatives have not yet produced detailed reporting requirements for NMR-based metabolomics experiments.

There are several online databases that allow deposition of NMR experiment results, such as NMRShiftDB for organic structures (http://www.nmrshiftdb.org/) and BioMagResBank (BMRB) (http://www.bmrb.wisc.edu/). These databases are mainly built to facilitate deposition of NMR spectra together with various amounts of associated metadata. In addition, data exchange formats for NMR data sets are available from both the CCPN project (Fogh et al., 2002; Vranken et al., 2005), that offers a data model for macromolecular NMR and related areas, and JCAMP-DX (Davies et al., 1993; Lampen et al., 1999). Also, a more general XML (eXtensible Markup Language) (http://www.w3.org/TR/xml11) format for analytical chemistry, that is currently in pre-release form, has been developed by the AnIML (Analytical Information Markup Language) (http://animl.sourceforge.net/) initiative. While these existing initiatives contain valuable content for handling metabolomics data sets, none have been developed based on a detailed systems analysis of an NMR-based metabolomics experiment.

The aim of this work was to perform a systems analysis of NMR-based metabolomics experiments in order to reveal their minimal reporting requirements. This will represent suggested core reporting requirements with the option of user-defined extra information. The results of such a systems analysis will not only enable the development of databases and data handling tools specifically for NMR-based metabolomics experiments, but will also enable proper assessment of the appropriateness of the pre-existing data models and data exchange formats for use in metabolomics data handling. We have produced an in-depth analysis of the NMR component of a metabolomics experiment, and finished the first draft of a data reporting standard. This has focussed on both 1D and 2D ^{1}H NMR experiments, but is also applicable to higher dimensions and other nuclei. We also report the modelling of this schema using Unified Modelling Language (UML) (Booch et al., 1999), and have extended this to a proof-of-concept implementation of the standard as an XML schema.

2. Scope of the proposed reporting requirements

A typical work flow for metabolomics experiments is depicted in Figure 1. This diagram divides the work flow into three major parts:

1. The source of sample material (experimental design; selection criteria for cell tissue or biofluid, cultivation or housing of biological source material; collection of samples from the biological source material and extraction of the metabolites that they contain).
2. The production of data sets (preparation of extracted samples for analysis by an analytical instrument; chemical analysis using a particular analytical technology; FID and spectral processing; spectral quantitation).
3. Statistical analysis and data mining of the data sets to provide answers to the original experimental questions.

INSERT FIGURE 1 HERE

The distinction between FID and spectral processing and spectral quantitation can be described as follows. FID and spectral processing is performed upon the raw output data from the analytical instrument. It involves transformation of a raw data set into a representation of the metabolome of the sample usually by mathematical or algorithmic means, i.e. production of an NMR spectrum from a FID (Free Induction Decay). Spectral quantitation is performed upon the data sets that result from FID and spectral processing and aims to summarise them or annotate them with speculative values by either automatic or manual means, e.g. techniques such as “bucketing” (also known as “binning”) and “peak-picking” perform spectral quantitation.

Of these activities, those involved in the production of data sets (contained within the dotted box within Figure 1) are dependent on the analytical technology that is used for chemical analysis, i.e. the choice of analytical technology will decide both how the extracted sample should be prepared for chemical analysis and the nature of the data sets that are produced and how they may be processed. The activities outside of the dotted box are not dependent on the analytical technology and may be performed in the same way regardless of the technology chosen for an experiment, i.e. an extracted sample may be divided and prepared separately for presentation to different analytical technologies and the algorithm underlying a data mining or statistical analysis technique will not change simply because the data to which it is applied is produced by a different instrument.

Working on the basis that this initiative would add to a number of pre-existing initiatives that aimed to provide data standards for metabolomics (Lindon et al., 2005; Bino et al., 2004; Jenkins et al., 2004), and in anticipation that it would become part of the recent community-led initiative to provide data standards not only for a range of analytical technologies but also for the complete metabolomics work flow (http://www.metabolomicssociety.org/mstandards.html), our systems analysis of NMR-based metabolomics focused on the analytical technology dependent activities involved in the production of data sets. The proposed reporting requirements aim to describe these activities and specify the references to data on the other activities in the metabolomics work flow that are needed to provide a complete description of an NMR-based metabolomics experiment.

A further decision regarding the scope of the systems analysis was to address both one- and two-dimensional NMR experiments. While at present the field of NMR-based metabolomics is dominated by the use of one-dimensional NMR methods, a number of recent studies have highlighted the significant value of two-dimensional NMR in metabolomics (Viant, 2003; Wang et al., 2003; Sandusky and Raftery, 2005). Therefore, by including two-dimensional experiments in the systems analysis we aim to increase the potential of the reporting requirements by enabling description of a new growth area in the field of NMR-based metabolomics. In theory, and because two-dimensional experiments and data are similar in structure to higher dimension experiments and data, we anticipate that these reporting requirements will also be appropriate for describing the third and higher dimensions of an NMR experiment; although higher dimensional experiments are not commonly performed in metabolomics at present.

3. Content of the proposed reporting requirements

Systems analysis of the activities involved in the production of data sets involves identification of a) the structure and content of the output data; and b) the set of meta-data items that describe how the output data was produced. The aim is to produce a complete data set that describes an experiment and its results in such a way that its output may be correctly interpreted and used by third parties. In Figure 2 we identify, for each activity involved in the production of data sets, the factors about which meta-data should be identified and the output data that must be analysed.

INSERT FIGURE 2 HERE

The following discussion of the content of the proposed reporting requirements will be structured according to the concepts provided in Figure 2. The meta-data items contained in the proposed reporting requirements are presented in Figure 3.

Meta-data

Sample Description

The sample description contains details of the biological sample, together with details of chemicals added to the sample to facilitate its analysis by NMR: one or more solvents added to the sample, chemicals added to the solvent to modify its properties (e.g. a buffer to alter the pH of the sample), an optional chemical shift standard used as an internal reference point for aligning spectra, optional internal standards for metabolite quantification and a field frequency compound to lock the spectrometer frequency. The sample description also contains a reference to information external to the reporting requirements which describes the history and provenance of the sample prior to its preparation for NMR analysis, i.e. a description of the activities involved in the production of sample material.

Analysis Description

For audit purposes, the reporting requirements specify that the date and time of data acquisition and contact details for the experimentalists responsible for the analysis should be recorded.

Instrument Description

An NMR instrument typically constitutes a number of components which may or may not have been constructed by the same manufacturer. As the type of instrument and, in particular, the software used for data acquisition can have an effect on the output data for a sample, the reporting requirements aim to capture this information in the instrument description.

Acquisition Parameters

Insight into the configuration of an analytical instrument at the time that a sample is analysed is crucial to correctly interpret the output data that is produced. The complete set of instrument parameters can be quite large and will contain a range of values from those that rarely change and have little impact on the output data to those that are highly variable and have a direct impact on the output data. The reporting requirements specify a small set of important parameters that should be recorded explicitly for each acquisition whilst at the same time requiring a reference to the acquisition parameters file produced by the acquisition software that contains the complete parameter set.

Of note within the acquisition parameters are values for the method of introducing the sample to the instrument and its size, e.g. 1mm tube, 50µl flow probe. These parameters are included to provide enough information to enable a judgement to be made about the actual quantity and dilution of sample material that has been analysed.

Quality Control

This type of information can be provided indirectly through description of the quality control procedures in place at the time of analysis of a sample, or directly via calculations performed on signals within the output data for the sample, e.g. a signal to noise ratio, the full width at half maximum (FWHM) of a reference peak or the width of a reference peak at 5% of its height. The reporting requirements specify that the latter two of these examples should be provided to enable third parties to make an assessment of the reliability of the data.

FID & Spectral Processing and Spectral Quantitation

All or part of FID and spectral processing is usually carried out under automation. As with acquisition parameters the full set of processing parameters will often be large and varied.

Here again the reporting requirements specify a small set of important parameters that should be recorded explicitly and also require a reference to the processing parameters file produced by the processing software that contains the complete parameter set (where available).

In general methods for FID and spectral processing are better defined and standardised than those for spectral quantitation. The reporting requirements reflect this through specification of much looser descriptive data items for the description of spectral quantitation..

INSERT FIGURE 3 HERE

Data Sets

The reporting requirements specify that at least one of the following should be provided for an analysis:

- A FID (or a reference to a file containing a FID)
- A spectrum that results from FID and spectral processing
- A spectrum that results from spectral quantitation

The reporting requirements specify the required content of FID and spectral based on the JCAMP-DX format for NMR (Davies et al., 1993; Lampen et al., 1999). This format was designed for spectral data transfer without loss of information. Use of the JCAMP-DX format during the systems analysis means that JCAMP-DX files may be used to fulfil the data sets part of the reporting requirement in those situations in which JCAMP-DX files are easily exported from an analytical instrument. The decision not to specify the use of JCAMP-DX explicitly means that the reporting requirements can also be fulfilled by other file formats where JCAMP-DX is not an export option, and future evolution of the reporting requirements, for example to support experiments with more than two dimensions, is not dependent on future evolution of JCAMP-DX.

The JCAMP-DX style for spectral representation involves specification of the units of measurement for the axes of a spectrum, the number of data points, and starting and ending values for the x-axis. The data matrix for a 1D spectrum can then be composed of either y-values alone (where the complete spectrum is being provided and the x-axis values can be calculated from the starting and ending values and number of data points) or (x,y) pairs which allows for the provision of only selected regions of a spectrum, which is a likely requirement for the reporting of NMR datasets by industry. For a 2D spectrum the data matrix is composed of a series of 1D spectra each annotated with a value for the second dimension. JCAMP-DX specifies a similar format for the representation of "peak-picked" spectra. The other common output of NMR spectral quantitation, the "bucketed" spectrum, is not supported by JCAMP-DX. Therefore, our reporting requirements specify the content for "bucketed" spectra following the JCAMP-DX style (see Figure 4).

INSERT FIGURE 4 HERE

4. Discussion

There are several pre-existing initiatives aimed at standardizing reporting requirements for both metabolomic experiments and for NMR based experiments. During our systems analysis these initiatives have been considered to ensure compatibility. In addition, our model has benefited from discussions with the wider metabolomics community, including academia and industry, at 'MetaboMeetings 1 and 2' in Cambridge, UK in 2005 and 2006 (http://smrsgroup.sourceforge.net/metabomeeting.html; http://www.mpdg.org/metabomeeting2/MM2_Program.htm).

Considering those initiatives that deal specifically with NMR-based data, there appears to be only a small amount of overlap between the reporting requirements and the schemata for pre-existing NMR spectral databases. The NMRShiftDB spectral library for organic structures and the BMRB database of quantitative data from NMR spectroscopic investigations both aim to provide resources of spectral information for the community to enable users to further their knowledge of biological systems. As such, they support information that is outside of the scope of our reporting requirements such as molecular descriptions to annotate spectra, whilst at the same time requiring fewer experimental meta-data.

However, there is considerable overlap between our proposed reporting requirements and the information captured in the data model produced by the CCPN initiative. Although CCPN is a low level description of the meta-data associated with an NMR experiment, and was originally designed to describe structural NMR experiments in enough detail to fully define a protein structure, the current compatibility suggests that converters could be produced to extract information from CCPN based databases to populate a minimal description of an NMR based metabolomics experiment that is based on the reporting requirements described here.

There is also substantial common ground between our proposed requirements and the 'Technique definition for NMR spectroscopy' described in AnIML. For example, most of the content of 'Measurement Parameters' in AnIML can be mapped one-to-one to the items in our 'Acquisition Parameters' section. The same applies to 'Processing Parameters' and

‘Instrument’ in AnIML, and ‘Post-Processing Parameters’ and ‘Instrument Description’ in our requirements. This implies mutual compatibility in terms of the information content and makes it possible to use AnIML as a format for exchange of data that complies with the reporting requirements without loss.

Similarly, and as mentioned above, JCAMP-DX for NMR may be used as a format for storing most of the data sets detailed in the reporting requirements. In addition, there is considerable overlap between the meta-data items that are listed in the reporting requirements and the JCAMP-DX header information and optional notes fields, making JCAMP-DX for NMR another format that may be used to exchange data sets that comply with parts of the requirements.

In terms of the current standardization documents related specifically to metabolomics experiments, our reporting requirements comply with the SMRS requirements for sample handling, data acquisition and instrument level data processing and FID and spectral processing. In this manner our project could be considered as being focused on a subset of the total SMRS description, which we have taken to a formal UML data model, as well as an XML-based implementation (see Current status and future development, below).

Considering previous initiatives within the plant metabolomics community, our reporting requirements also comply with the MIAMET recommendations that resulted from discussions at the International Plant Metabolomics Congress in April 2002 and April 2003 (http://www.metabolomics-2003.mpg.de/) and their organisation means that they may readily be implemented as sub-components of an ArMet core implementation, thereby placing them within the context of a complete metabolomics experiment.

Finally, as well as examining compatibility with previous initiatives focussed on describing metabolomic experiments or NMR spectroscopy data, it is important to consider how this work fits in with the wider functional genomic world. FuGE offers a model of the shared components in different functional genomics domains, such as experimental design, sample preparation, subject selection criteria, etc. FuGE has attracted substantial support in the standard development community as a possible common ground for integration of various functional genomics data standards. It has been adopted by MGED and PSI and is under consideration by the MSI. In this respect, we have developed our description so as to be compatible with the wider FuGE description and, following discussions with the FuGE team, they have created a proof-of-concept implementation of the reporting requirements using the current version of FuGE (Andy Jones, personal communication).

5. Current status and future development

Our main aim in this project was to generate a proposal for reporting requirements based on a systems analysis of an NMR-based metabolomics experiment. This has led to the design of a UML object model as a proof-of-concept. The development of a data model was a natural next step in the systems analysis and enabled us to place the descriptions in a more formal context and identify potential problems with its implementation. It has allowed us to identify objects in the domain and specify relationships between them as well as set restrictions on their attributes. Further formalization of this model has been done in order to implement the UML object model as an XML schema. The full reporting requirements and our example object model are available as Electronic Supplementary Material to the article. Work on these reporting requirements was started following “MetaboMeeting 1”, Cambridge, 2005, and prior to the creation of the MSI. We anticipate that development of these requirements will be taken forward by the MSI; specifically by the Chemical

Analysis Working Group on which the authors have representation. Terminology from the requirements has already been provided as an input to the Ontologies working group on which we also have representation. It is hoped that under the auspices of the MSI these reporting requirements may be refined and improved to produce a data standard that is of use to the metabolomics community as a whole.

References

Bino, R. J., Hall, R. D., Fiehn, O., Kopka, J., Saito, K., Draper, J., Nikolau, B. J., Mendes, P., Roessner-Tunali, U., Beale, M. H., Trethewey, R. N. , Lange, B. M. , Wurtele, E. S. and Sumner, L. W. (2004) Potential of metabolomics as a functional genomics tool. Trends Plant Sci. 9, no. 9, 418-425.

Booch, G., Rumbaugh, J., Jacobson, I. (1999) The Unified Modeling Language User Guide, Addison-Wesley, Reading, MA.

Davies, A.N. and Lampen, P. (1993) JCAMP-DX for NMR. Applied Spectroscopy, 47 (8), 1093-1099

Fogh, R., Ionides, J., Ulrich, E., Boucher, W., Vranken, W., Linge, J. P., Habeck, M., Rieping, W., Bhat, T. N., Westbrook, J., et al. (2002) The CCPN project: an interim report on a data model for the NMR community. Nat Struct Biol 9, 416-418.

Griffin, J. L., and Shockcor, J. P. (2004) Metabolic profiles of cancer cells. Nat Rev Cancer 4, 551-561.

Griffin, J. L., Williams, H. J., Sang, E., Clarke, K., Rae, C., and Nicholson, J. K. (2001) Metabolic profiling of genetic disorders: a multitissue (1)H nuclear magnetic resonance spectroscopic and pattern recognition study into dystrophic tissue. Anal Biochem 293, 16-21.

Jenkins, H., Hardy, N., Beckmann, M., Draper, J., Smith, A. R., Taylor, J., Fiehn, O., Goodacre, R., Bino, R. J., Hall, R., et al. (2004) A proposed framework for the description of plant metabolomics experiments and their results. Nat Biotechnol 22, 1601-1606.

Lampen, P., Lambert, J., Lancashire, R.J., McDonald, R.S., McIntyre, P.S., Rutledge, D.N., Fröhlich, T. and Davies, A.N. (1999) An extension to the JCAMP-DX standard file format, JCAMP-DX V.5.01. Pure and Applied Chemistry, 7(8), 1549-1556.

Lindon, J. C., Nicholson, J. K., Holmes, E., Antti, H., Bollard, M. E., Keun, H., Beckonert, O., Ebbels, T. M., Reily, M. D., Robertson, D., et al. (2003) Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project. Toxicol Appl Pharmacol 187, 137-146.

Lindon, J. C., Holmes, E., Nicholson, J. K. (2004) Toxicological applications of magnetic resonance. Progress in NMR Spectroscopy 45, 109-143.

Lindon, J. C., Nicholson, J. K., Holmes, E., Keun, H. C., Craig, A., Pearce, J. T., Bruce, S. J., Hardy, N., Sansone, S. A., Antti, H., et al. (2005) Summary recommendations for standardization and reporting of metabolic analyses. Nat Biotechnol 23, 833-838.

Nicholson, J. K., Connelly, J., Lindon, J. C., and Holmes, E. (2002) Metabonomics: a platform for studying drug toxicity and gene function. Nat Rev Drug Discov 1, 153-161.

Nicholson, J. K., Holmes, E., and Wilson, I. D. (2005). Gut microorganisms, mammalian metabolism and personalized health care. Nat Rev Microbiol 3, 431-438.

Sandusky, P., and Raftery, D. (2005) Use of selective TOCSY NMR experiments for quantifying minor components in complex mixtures: Application to the metabonomics of amino acids in honey. Anal Chem 77, 2455-2463.

Spellman, P. T., Miller, M., Stewart, J., Troup, C., Sarkans, U., Chervitz, S., Bernhart, D., Sherlock, G., Ball, C., Lepage, M., et al. (2002) Design and implementation of microarray gene expression markup language (MAGE-ML) Genome Biol 3, RESEARCH0046.

Viant, M. R., Rosenblum, E. S., and Tjeerdema, R. S. (2003) NMR-based metabolomics: A powerful approach for characterizing the effects of environmental stressors on organism health. Env Sci Technol 37, 4982-4989.

Viant, M. R. (2003) Improved methods for the acquisition and interpretation of NMR metabolomic data. Biochem Biophys Res Comm 10, 943-948.

Vranken, W. F., Boucher, W., Stevens, T. J., Fogh, R. H., Pajon, A., Llinas, M., Ulrich, E. L., Markley, J. L., Ionides, J., and Laue, E. D. (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins 59, 687-696.

Wang, Y. L., Bollard, M. E., Keun, H., Antti, H., Beckonert, O., Ebbels, T. M., Lindon, J. C., Holmes, E., Tang, H. R., and Nicholson, J. K. (2003) Spectral editing and pattern recognition methods applied to high-resolution magic-angle spinning H-1 nuclear magnetic resonance spectroscopy of liver tissues. Anal Biochem 323, 26-32.

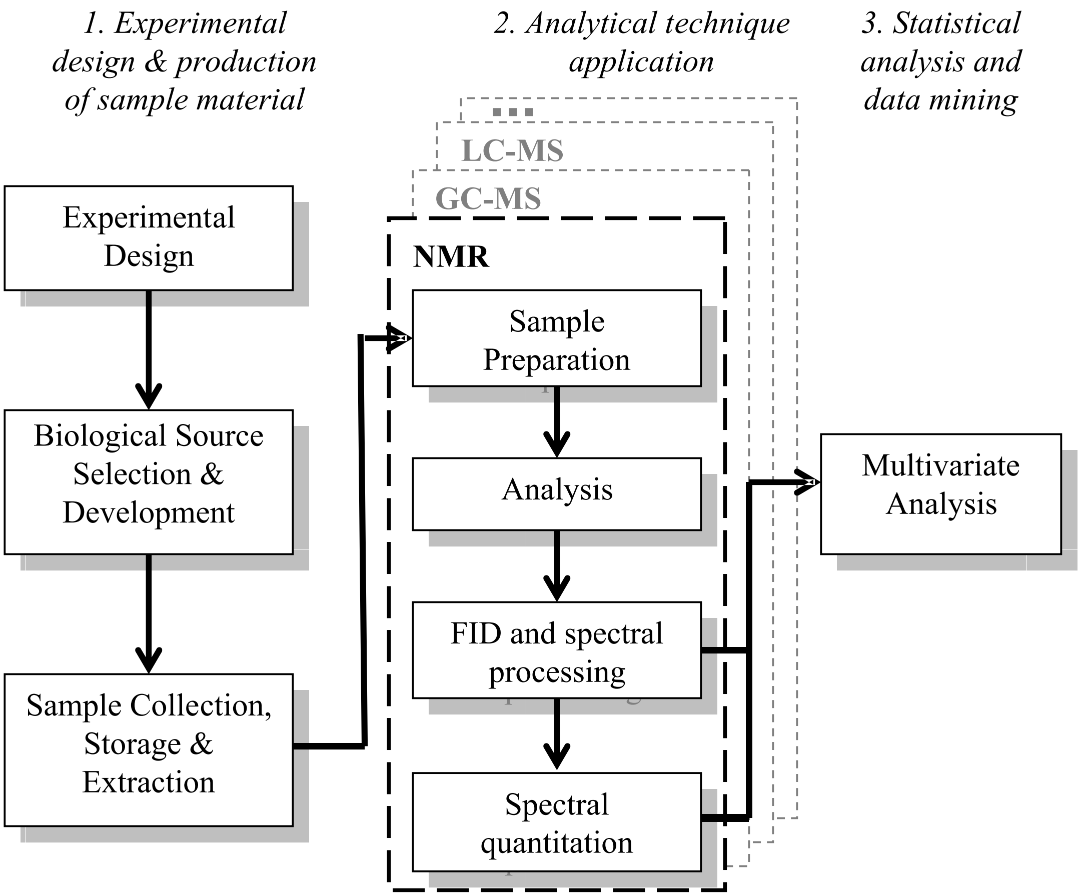

1. Experimental design & production of sample material
2. Analytical technique application
3. Statistical analysis and data mining
...
LC-MS
GC-MS
NMR
Experimental Design
Biological Source Selection & Development
Sample Collection, Storage & Extraction
Sample Preparation
Analysis
FID and spectral processing
Spectral quantitation
Multivariate Analysis

Figure 1: The work flow of a metabolomics experiment

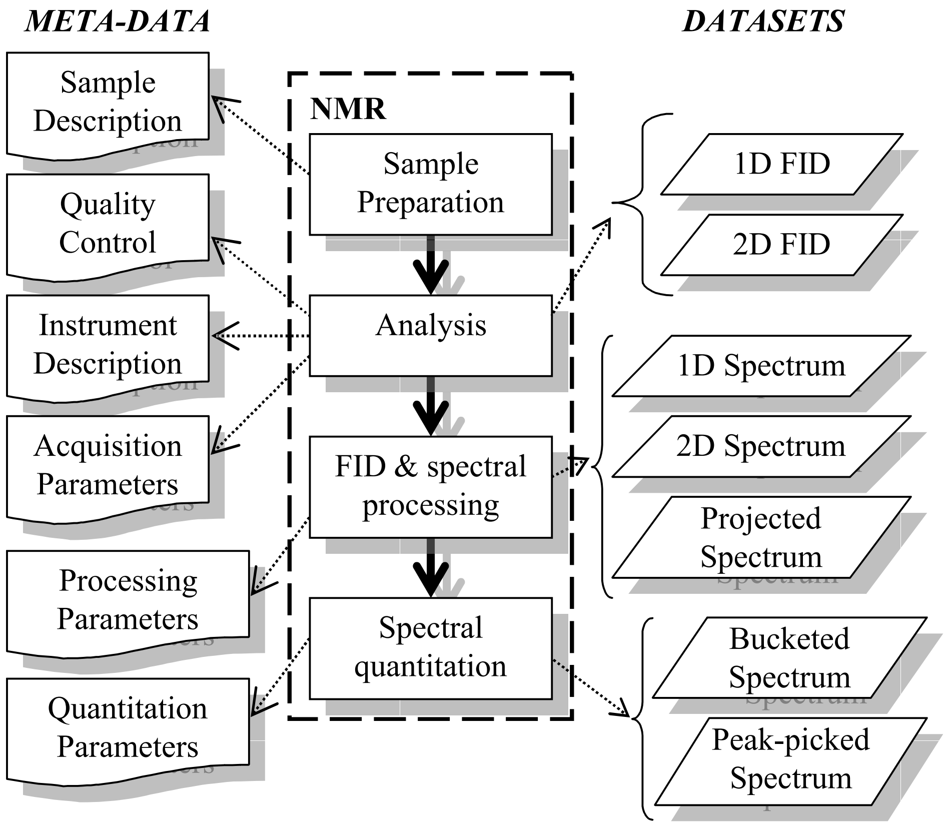
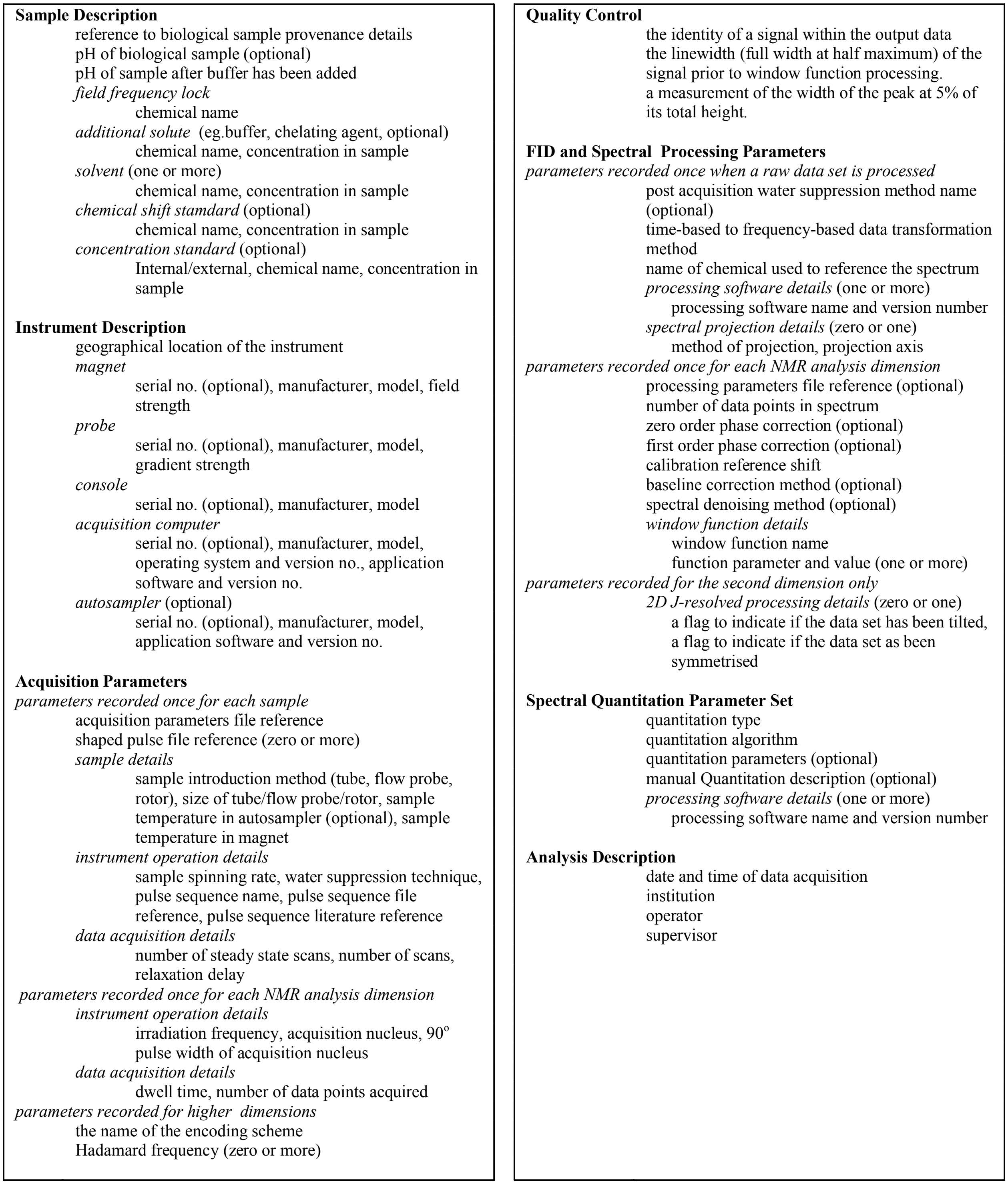


Figure 2: NMR data set production, meta-data and output data

Figure 3 Revised

Click here to download Figure: Fig3.ppt



Sample Description

reference to biological sample provenance details
pH of biological sample (optional)
pH of sample after buffer has been added
field frequency lock
chemical name
additional solute (eg.buffer, chelating agent, optional)
chemical name, concentration in sample
solvent (one or more)
chemical name, concentration in sample
chemical shift stamdard (optional)
chemical name, concentration in sample
concentration standard (optional)
Internal/external, chemical name, concentration in sample

Instrument Description

geographical location of the instrument
magnet
serial no. (optional), manufacturer, model, field strength
probe
serial no. (optional), manufacturer, model, gradient strength
console
serial no. (optional), manufacturer, model
acquisition computer
serial no. (optional), manufacturer, model, operating system and version no., application software and version no.
autosampler (optional)
serial no. (optional), manufacturer, model, application software and version no.

Acquisition Parameters

parameters recorded once for each sample
acquisition parameters file reference
shaped pulse file reference (zero or more)
sample details
sample introduction method (tube, flow probe, rotor), size of tube/flow probe/rotor, sample temperature in autosampler (optional), sample temperature in magnet
instrument operation details
sample spinning rate, water suppression technique, pulse sequence name, pulse sequence file reference, pulse sequence literature reference
data acquisition details
number of steady state scans, number of scans, relaxation delay
parameters recorded once for each NMR analysis dimension
instrument operation details
irradiation frequency, acquisition nucleus, 90^{o} pulse width of acquisition nucleus
data acquisition details
dwell time, number of data points acquired
parameters recorded for higher dimensions
the name of the encoding scheme
Hadamard frequency (zero or more)

Quality Control

the identity of a signal within the output data
the linewidth (full width at half maximum) of the signal prior to window function processing.
a measurement of the width of the peak at 5% of its total height.

FID and Spectral Processing Parameters

parameters recorded once when a raw data set is processed
post acquisition water suppression method name (optional)
time-based to frequency-based data transformation method
name of chemical used to reference the spectrum
processing software details (one or more)
processing software name and version number
spectral projection details (zero or one)
method of projection, projection axis
parameters recorded once for each NMR analysis dimension
processing parameters file reference (optional)
number of data points in spectrum
zero order phase correction (optional)
first order phase correction (optional)
calibration reference shift
baseline correction method (optional)
spectral denoising method (optional)
window function details
window function name
function parameter and value (one or more)
parameters recorded for the second dimension only
2D J-resolved processing details (zero or one)
a flag to indicate if the data set has been tilted,
a flag to indicate if the data set as been symmetrised

Spectral Quantitation Parameter Set

quantitation type
quantitation algorithm
quantitation parameters (optional)
manual Quantitation description (optional)
processing software details (one or more)
processing software name and version number

Analysis Description

date and time of data acquisition
institution
operator
supervisor

Figure 3: Meta-data items specified in the proposed reporting requirements (items in bold and italics represent groups and sub-groups of data items respectively whilst the data items themselves are in normal font)

Bucketed Spectra

units of measurement on the x-axis, y-axis
number of buckets in the spectrum.
data matrix for the spectrum (the starting x-axis value for the bucket, the ending x-axis value for the bucket, the x-axis value at the centre of the bucket, the y-axis value)

Peak-picked Spectra

units of measurement on the x-axis, y-axis
number of peaks in the spectrum.
data matrix for the spectrum (the x-axis value for the peak, the y-axis value for the peak, the type of the peak)

Figure 4: The content for "bucketed" and "peak-picked" spectra items specified in the proposed reporting requirements (items in bold and italics represent groups and sub-groups of data items respectively whilst the data items themselves are in normal font)

Proposed Reporting Requirements for NMR-based Metabolomics

Helen Jenkins
Nigel Hardy
Denis V. Rubtsov
Julian L. Griffin
Mark R. Viant
Christian Ludwig
Ulrich Guenther
John Easton

$Revision: 1.4 $

Table of Contents

1. Status of this document

This document is the outcome of three meetings involving Mark Viant, Christian Ludwig, John Easton and Ulrich Guenther from Birmingham, Denis Rubtsov from Cambridge and Helen Jenkins and Nigel Hardy (3rd meeting only) from Aberystwyth that were held at the Henry Wellcome Building for Biomolecular NMR Spectroscopy on the 25th of July 2005, the 18th of August 2005 and the 1st of December 2005. The aim of the first two meetings was to compare and merge the concepts (and their descriptions) encapsulated in the Cambridge and Aberystwyth (ArMet-compliant) NMR data models and the instrument parameters and data processing parameters models developed at Birmingham. The Cambridge model is based on experience from both Cambridge and Imperial College London (based upon the SMRS policy document), whilst the Aberystwyth model is based on requirements for the UK Centre for Plant and Microbial Metabolomics based at Rothamsted Research. The aim of the third meeting was to compare and devise a merging strategy for the two XML implementations of the standard that were produced at Aberystwyth and Cambridge.

Presentation of the proposed standard at MetaboMeeting2.0 in Cambridge on the 10th of January, 2006 elicited positive feedback from the community. This document represents the results of re-working the original document to produce a standalone representation of the standard suitable for distribution to interested parties for more detailed assessment.

2. Concepts addressed

The tables in this section represent the concepts that were discussed at the meetings and contain the data items that are required to describe them. A *Block* is a distinct subset of the data items for a concept. The *Domain* column indicates the data types for the data items. The *Units* column contains candidate units of measurement for numeric data items. The column headed *?* contains an indication of whether a data item is a required (*R*) or optional (*O*) element of its Block.

Following each table is a list of the association rules that describe how the blocks of data items for the concept are related. There is also a UML representation of the structure of the concept.

2.1. Analysis

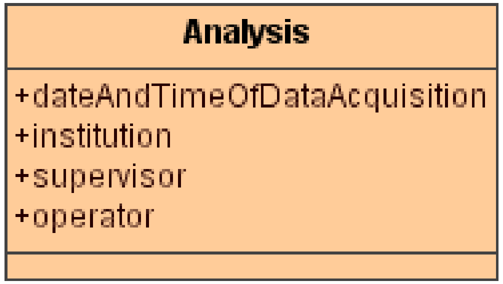
In this context an *Analysis* is NMR data acquisition for a single sample.

Block	**Field**				
	Name	**Definition**	**Domain**	**Units**	**?**
Analysis					
	dateAndTimeOfDataAcquisition	The date/time at which the analysis of a sample by NMR was performed.	ISO 8601		R

Block	**Field**				
	Name	**Definition**	**Domain**	**Units**	**?**
	institution	The name of the institution at which the analysis was carried out.	string		R
	supervisor	The name of the supervisor who oversaw the analysis. by NMR.	string		R
	operator	The name of the operator who carried out the analysis.	string		R

Association rules. As there is only one data block for Analysis there are no association rules.

Figure 1. UML for the Analysis concept



2.2. Sample

Block	**Field**				
	Name	**Definition**	**Domain**	**Units**	**?**
NMR Sample					
	originalBiologicalSampleReference	A reference to information on the provenance of the original biological source material.	URI		R
	originalBiologicalSamplepH	The pH value of the original biological sample material.	float	pH	O
	postBufferpH	The pH value of the sample after the buffer has been added.	float	pH	R
	concentrationOfSoluteInSample	The concentration of additional solute (e.g. buffer, chelating agent) within the sample.	float	('moles' | 'millimoles')	R
	concentrationOfChemShiftStdInSample	The concentration of chemical shift standard within the sample.	float	('moles' | 'millimoles')	O

Block	**Field**				
	Name	**Definition**	**Domain**	**Units**	**?**
	concentrationOfSolventInSample	The concentration of a solvent within the sample.	float	('moles' | 'millimoles')	R
	concentrationOfConcentrationStdInSample	The concentration of concentration standard compound within the sample.	float	('moles' | 'millimoles')	R
	concentrationStdType	An indication of the type of the concentration standard within the sample.	See §4.1		R
Field Frequency Lock					
	fieldFrequencyLockName	The name of a field frequency lock compound.	string		R
Additional Solute					
	soluteName	The name of a solute that is added to a sample, e.g. a buffer or chelating agent.	string		R
Chemical Shift Standard					
	chemicalShiftStdName	The name of a compound added to a sample to enable alignment of spectra.	string		R
Solvent					
	solventName	The name of a solvent.	string		R
Concentration Standard					
	concentrationStdName	The name of a concentration standard compound.	string		R

Association rules. The following association rules apply to the data blocks for the Sample concept:

- An NMR Sample is associated with one Field Frequency Lock
- An NMR Sample is associated with one Additional Solute
- An NMR Sample is optionally associated with one Chemical Shift Standard
- An NMR Sample is associated with at least one and possibly many Solvents
- An NMR Sample is associated with one Concentration Standard.

2.3. Instrument

Block	**Field**				
	Name	**Definition**	**Domain**	**Units**	**?**
Instrument					
	location	The geographical location of the instrument	string		R
Magnet					

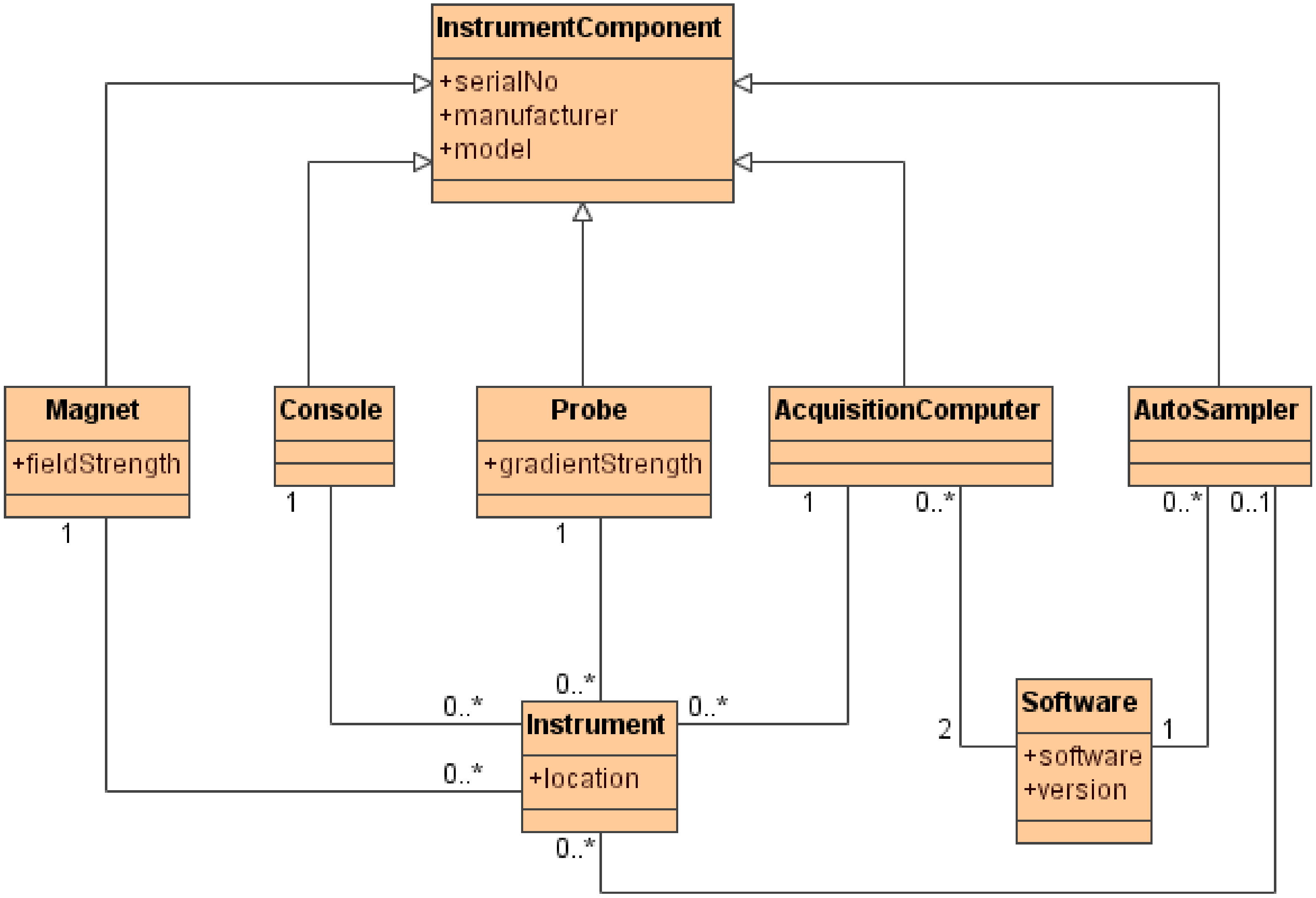
Block	**Field**				
	Name	**Definition**	**Domain**	**Units**	**?**
	serialNo	The unique serial number for an NMR magnet.	string		O
	manufacturer	The manufacturer of an NMR magnet.	string		R
	model	The manufacturer's model identifier for an NMR magnet.	string		R
	fieldStrength	The strength of the magnetic field produced by an NMR magnet.	float	('gauss' | 'tesla')	R
Probe					
	serialNo	The unique serial number for an NMR probe.	string		O
	manufacturer	The manufacturer of an NMR probe (may be "custom made")	string		R
	model	The manufacturer's model identifier for an NMR probe.	string		R
	gradientStrength	The variation in magnetic field between the gradient coils in an NMR probe.	float	('gauss' | 'tesla')	O
Console					
	serialNo	The unique serial number for an NMR console.	string		O
	manufacturer	The manufacturer of an NMR console.	string		R
	model	The manufacturer's model identifier for an NMR console.	string		R
Acquisition Computer					
	serialNo	The unique serial for an NMR acquisition computer.	string		O
	manufacturer	The manufacturer of an NMR acquisition computer.	string		R
	model	The manufacturer's model identifier for an NMR acquisition computer.	string		R
	operatingSystemSoftware	The name of the operating system on an NMR acquisition computer.	string		R
	operatingSystemVersion	The version of the operating system on an NMR acquisition computer.	string		R
	applicationSoftware	The name of the software used for acquisition on an NMR acquisition computer.	string		R
	applicationSoftwareVersion	The version of the acquisition software on an NMR acquisition computer.	string		R
Autosampler					
	serialNo	The unique serial number for an NMR autosampler.	string		O
	manufacturer	The manufacturer of an NMR autosampler.	string		R
	model	The manufacturer's model identifier for an NMR autosampler.	string		R

Block	Field				
	Name	**Definition**	**Domain**	**Units**	**?**
	applicationSoftware	The name of software used to control an NMR autosampler.	string		R
	applicationSoftwareVersion	The version of the software used to control an NMR autosampler.	string		R

Association rules. The following association rules apply to the data blocks for the Instrument concept:

- An Instrument is associated with one Magnet
- An Instrument is associated with one Probe
- An Instrument is associated with one Console
- An Instrument is associated with one Acquisition Computer
- An Instrument is optionally associated with one Autosampler

Figure 3. UML for the Instrument concept



2.4. Instrument Acquisition Parameters

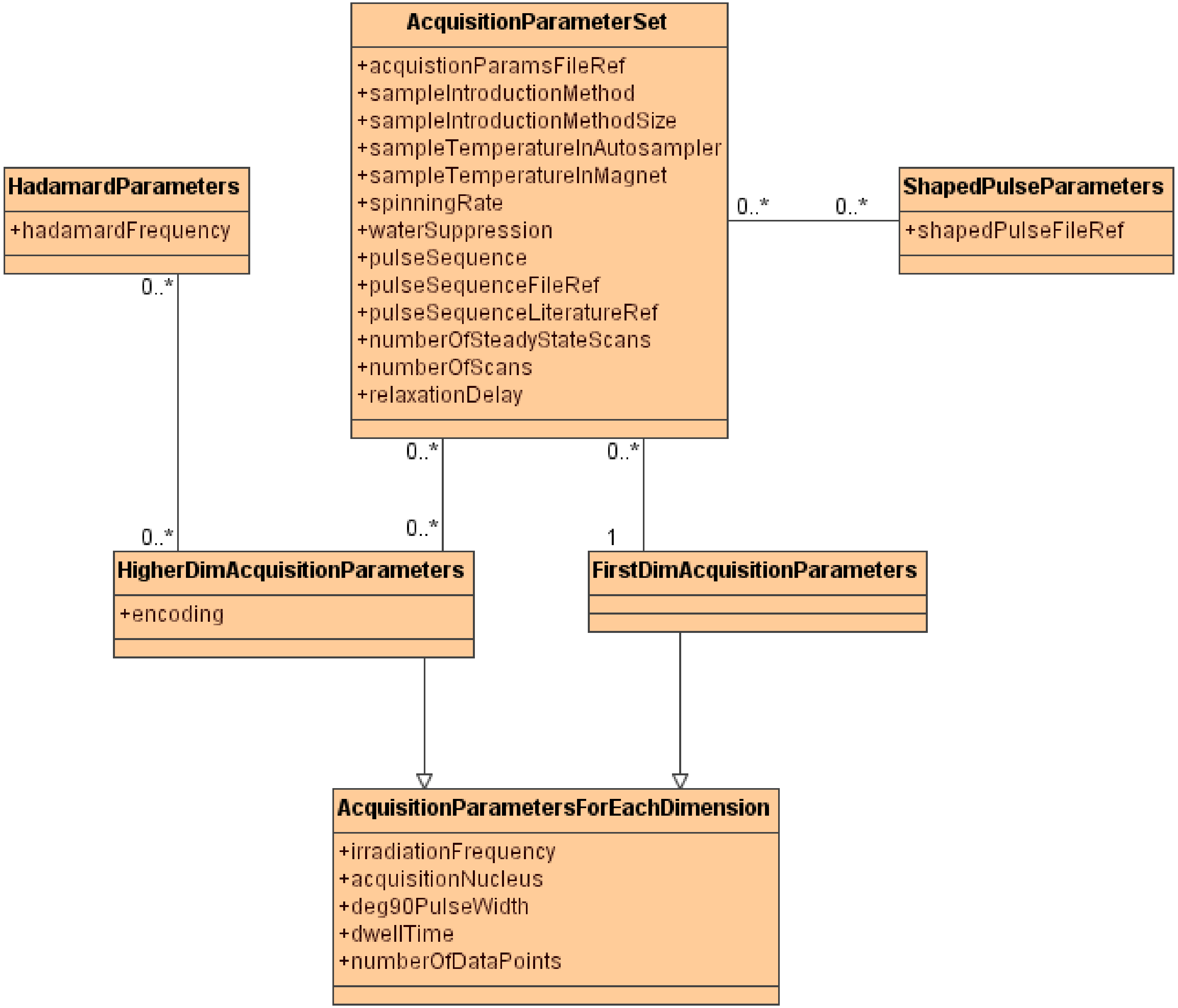
Block	**Field**				
	Name	**Definition**	**Domain**	**Units**	**?**
Acquisition Parameter Set					
	acquisitionParamsFileRef	A reference to the file of acquisition parameters produced by the instrument	URI		R
	sampleIntroductionMethod	The method of introduction of the sample to the spectrometer	See §4.2		R
	sampleIntroductionMethodSize	The size of the tube or rotor or the active volume of the flow probe	float	('millilitres' | 'microlitres' | 'millimetres')	R
	sampleTemperatureInAutosampler	The temperature of the sample whilst in the autosampler if fitted	float	('centigrade' | 'kelvin' | 'fahrenheit')	O
	sampleTemperatureInMagnet	The temperature of the sample whilst in the magnet	float	('centigrade' | 'kelvin' | 'fahrenheit')	R
	spinningRate	The rate at which the sample is spun to improve resolution by partially averaging out inhomogeneities in the magnetic field	float	('hertz' | 'kilohertz' | 'megahertz')	R
	waterSuppression	The technique used to suppress the water peak in the spectrum	See §4.3		R
	pulseSequence	The pulse sequence name	See §4.4		R
	pulseSequenceFileRef	A reference to a file that specifies the pulse sequence	URI		R
	pulseSequenceLiteratureRef	A reference to a description of the pulse sequence in the literature	URI		R
	numberOfSteadyStateScans	The number of scans whose data is not summed to create the spectrum for a sample, but that are carried out to establish the steady-state of relaxation for the nuclei	integer		R
	numberOfScans	The number of repeat scans to be performed and summed to create the spectrum for a sample	integer		R
	relaxationDelay	The delay between repeat scans to allow the nuclei to relax back to their steady-state	float	('seconds' | 'milliseconds' | 'microseconds')	R
Acquisition Parameters Recorded For Each Dimension					
	irradiationFrequency	The frequency of RF radiation used to irradiate a sample	float	('hertz' | 'kilohertz' | 'megahertz')	R
	acquisitionNucleus	The nucleus being studied	string		R
	deg90PulseWidth	The 90 degree pulse width of acquisition nucleus	float	('seconds' | 'milliseconds' | 'microseconds')	R
	dwellTime	The digital sampling interval	float	('seconds' | 'milliseconds' | 'micro-	R

Block	**Field**				
	Name	**Definition**	**Domain**	**Units**	**?**
				seconds')	
	noOfDataPoints	The number of data points acquired (should match the number of datapoints in the FID dataset(s) when describing the first dimension, or the additional axis of the 2D FID when describing the second dimension).	integer		R
Acquisition Parameters Recorded For Second and Higher Dimensions					
	encoding	The scheme for producing a numerical representation of the environment of an atom during an NMR experiment	See §4.5		R
Shaped Pulse Parameters					
	shapedPulseFileRef	A reference to a file containing a specification of the shape of an excitation pulse	URI		R
Hadamard Parameters					
	hadamardFrequency	A Hadamard frequency used during Hadamard encoding	float	('hertz' | 'kilohertz' | 'megahertz')	R

Association rules. The following association rules apply to the data blocks for the Instrument Acquisition Parameters concept:

- When describing a 1D NMR experiment an Acquisition Parameter Set is associated with:
  - One set of Acquisition Parameters Recorded for Each Dimension
  - Zero or many sets of Shaped Pulse Parameters
- When describing an NMR experiment with 2 or more dimensions an Acquisition Parameter Set is associated with:
  - A set of Acquisition Parameters Recorded for Each Dimension for the first dimension and for each additional dimension.
  - A set of Acquisition Parameters Recorded for the Second and Higher Dimensions for each additional dimension each of which is associated with:
    - One or more sets of Hadamard Parameters if, and only if, the second dimension *encoding* data item contains the value *Hadamard*
  - Zero or many sets of Shaped Pulse Parameters

Figure 4. UML for the Instrument Acquisition Parameters concept

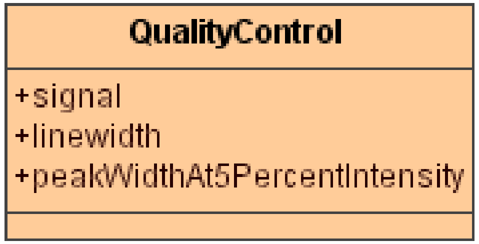

AcquisitionParameterSet
+acquistionParamsFileRef
+sampleIntroductionMethod
+sampleIntroductionMethodSize
+sampleTemperatureInAutosampler
+sampleTemperatureInMagnet
+spinningRate
+waterSuppression
+pulseSequence
+pulseSequenceFileRef
+pulseSequenceLiteratureRef
+numberOfSteadyStateScans
+numberOfScans
+relaxationDelay
HadamardParameters
+hadamardFrequency
ShapedPulseParameters
+shapedPulseFileRef
0..*
0..*
0..*
0..*
0..*
0..*
0..*
1
HigherDimAcquisitionParameters
+encoding
FirstDimAcquisitionParameters
AcquisitionParametersForEachDimension
+irradiationFrequency
+acquisitionNucleus
+deg90PulseWidth
+dwellTime
+numberOfDataPoints

2.5. Quality Control

Block	**Field**				
	Name	**Definition**	**Domain**	**Units**	**?**
Quality Control					
	signal	The identity of signal used for checking	string		R
	linewidth	The linewidth (FWHM: full width at half maximum) measured for the chosen signal in the 1D data in absence of Window Function processing	float	('hertz' | 'kilohertz' | 'megahertz')	R
	peak-WidthAt5PercentIntensity	A measurement of the width of the peak at 5% of its total height.	float	percentage	R

Association rules. As there is only one data block for Quality Control there are no association rules.

Figure 5. UML for the Quality Control concept



2.6. Data Processing

Data processing has been split into two parts. This split is based on the datasets that are produced as a result of different levels of data processing (described below). 1D FID and multi-dimension FID data are as produced by the instrument. 1D Spectra, mutli-dimension Spectra and Projected Spectra are produced by FID and spectral processing (spectra generation from FID data). Bucketed Spectra and Peak-Picked Spectra are produced by spectral quantitation.

Block	**Field**				
	Name	**Definition**	**Domain**	**Units**	**?**
FID & Spectral Processing Parameter Set					
	postAcquisitionWaterSuppression	The data processing technique used to suppress the water peak in the spectrum.	See §4.6		O
	transformationType	The method use to transform the time-based acquisition data into frequency-based	See §4.7		R

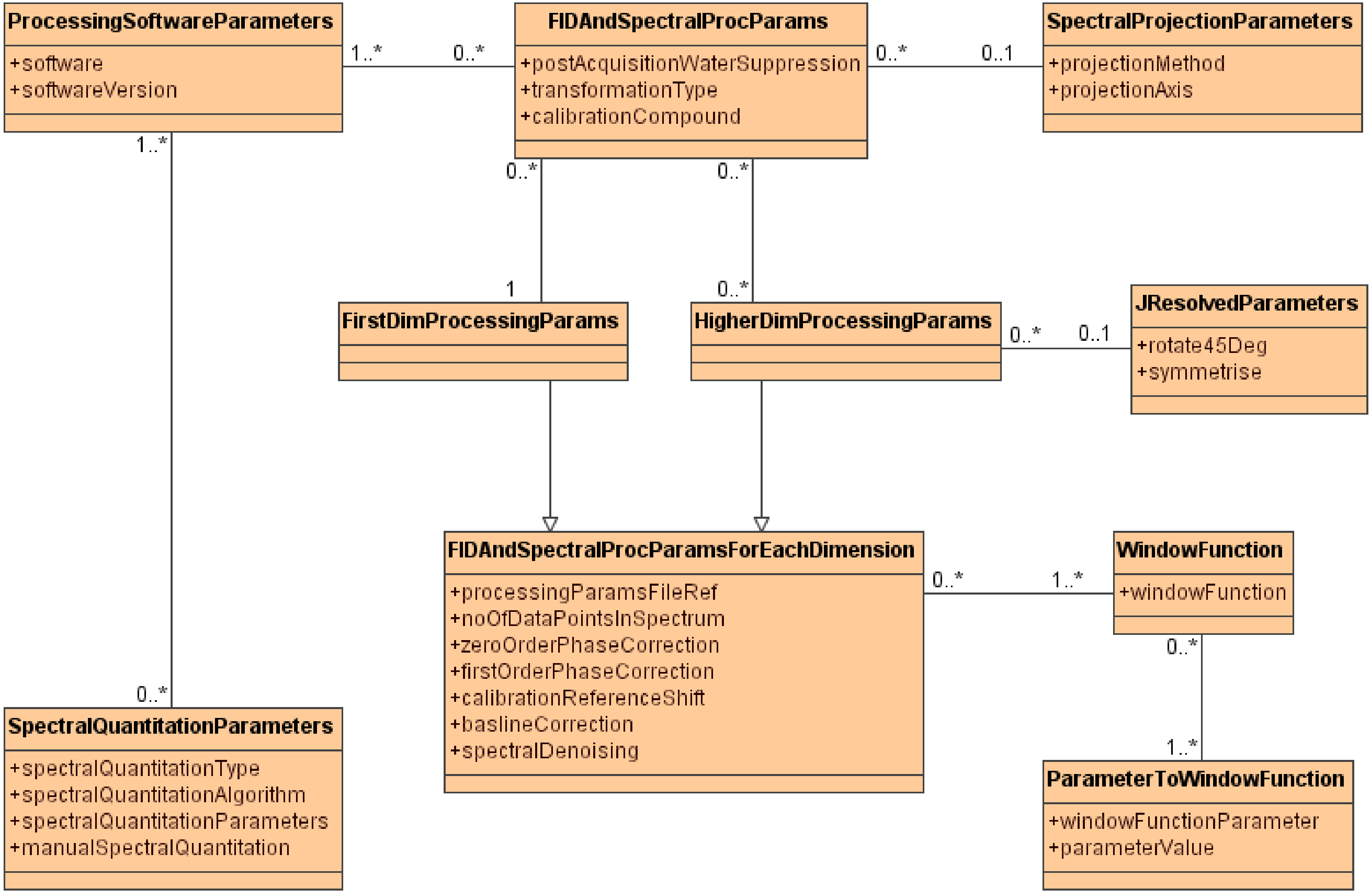
Block	Field				
	Name	Definition	Domain	Units	?
		data.			
	calibrationCompound	The chemical identity used to reference the spectrum (default to Chemical Shift Standard where available)	string		R
Processing Parameters Recorded for Each Dimension					
	processingParamsFileRef	A reference to the file of processing parameters produced by the instrument	URI		O
	noOfDataPointsInSpectrum	The number of data points in the spectrum that results from data pre-processing (should match the number of datapoints in the spectrum when describing the first dimension, or the number of data points in the additional axis of the 2D spectrum when describing the second dimension)	integer		R
	zeroOrderPhaseCorrection	The number of degrees of the zero order phase adjustment.	float	degrees	O
	firstOrderPhaseCorrection	The number of degrees of the first order phase adjustment.	float	degrees	O
	calibrationReferenceShift	The parts-per-million value of the peak used to reference the spectrum.	integer	ppm	R
	baselineCorrection	A description of the approach to flattening the baseline of the spectrum that results from data pre-processing.	string		O
	spectralDenoising	A description of any processing carried out to eliminate or reduce the noise in a spectrum.	string		O
Window Function Parameters					
	windowFunction	A function applied to a FID to increase the SNR or the resolution.	See §4.8		R
Parameter to Window Function Parameters					
	windowFunctionParameter	The name of a parameter to a Window Function	See §4.9		R
	parameterValue	The value for a parameter to a Window Function	string		R
2D J-Resolved Processing Parameters					
	rotate45Deg	An indication of whether a data set resulting from 2D J-Resolved analysis was rotated as part of data pre-processing.	Boolean		R
	symmetrise	An indication of whether a data set resulting from 2D J-Resolved analysis was symmetrised about the horizontal axis as part of data pre-processing.	Boolean		R
Processing Software Parameters					
	software	The name of a software artifact used during data processing.	string		R
	softwareVersion	The version of a software artifact used during data processing.	string		R
Spectral Projection Parameters					
	projectionMethod	A method of spectral projection.	See §4.10		R

Block	**Field**				
	Name	**Definition**	**Domain**	**Units**	**?**
	projectionAxis	The axis onto which a 2D spectrum was projected.	See §4.11		R
Spectral Quantitation Parameter Set					
	spectralQuantitationType	The approach to spectral quantitation.	See §4.12		R
	spectralQuantitationAlgorithm	A description of the approach to spectral quantitation.	string		R
	spectralQuantitationParameters	A description of the parameters used to govern spectral quantitation.	string		O
	manualSpectralQuantitation	A description of any manual manipulation or tidying of the data performed during spectral quantitation.	string		O

Association rules. The following association rules apply to the data blocks for the Data Processing concept:

- When describing FID and spectral processing in a 1D NMR experiment a FID & Spectral Processing Parameter Set is associated with:
  - One set of Processing Parameters Recorded for Each Dimension which is associated with:
    - One or more sets of Window Function Parameters each of which is associated with:
      - One of more sets of Parameter to Window Function Parameters
  - One or more sets of Processing Software Parameters
- When describing multi-dimensional FID and spectral processing without spectral projection a FID & Spectral Processing Parameter Set is associated with:
  - One set of Processing Parameters Recorded for Each Dimension to describe the first dimension which is associated with:
    - One or more sets of Window Function Parameters each of which is associated with:
      - One of more sets of Parameter to Window Function Parameters
  - A set of Processing Parameters Recorded for Each Dimension to describe each additional dimension each of which is associated with:
    - One or more sets of Window Function Parameters each of which is associated with:
      - One of more sets of Parameter to Window Function Parameters
    - Zero or one set of 2D J-Resolved Processing Parameters
  - One or more sets of Processing Software Parameters
- When describing multi-dimensional FID and spectral processing with spectral projection a FID & Spectral Processing Parameter Set is associated with:
  - One set of Processing Parameters Recorded for Each Dimension to describe the first dimension which is associated with:
    - One or more sets of Window Function Parameters each of which is associated with:
      - One of more sets of Parameter to Window Function Parameters
  - A set of Processing Parameters Recorded for Each Dimension to describe each additional dimension each of which is associated with:
    - One or more sets of Window Function Parameters each of which is associated with:
      - One of more sets of Parameter to Window Function Parameters
    - Zero or one set of 2D J-Resolved Processing Parameters
  - One or more sets of Processing Software Parameters
  - One set of Spectral Projection Parameters
- When describing spectral quantitation a Spectral Quantitation Parameter Set is associated with:
  - One or more sets of Processing Software Parameters

Figure 6. UML for the Data Processing concept


ProcessingSoftwareParameters
+software
+softwareVersion
1..*
0..*
FIDAndSpectralProcParams
+postAcquisitionWaterSuppression
+transformationType
+calibrationCompound
0..*
0..1
SpectralProjectionParameters
+projectionMethod
+projectionAxis
1..*
0..*
0..*
1
0..*
FirstDimProcessingParams
HigherDimProcessingParams
0..*
0..1
JResolvedParameters
+rotate45Deg
+symmetrise
FIDAndSpectralProcParamsForEachDimension
+processingParamsFileRef
+noOfDataPointsInSpectrum
+zeroOrderPhaseCorrection
+firstOrderPhaseCorrection
+calibrationReferenceShift
+baslineCorrection
+spectralDenoising
0..*
1..*
WindowFunction
+windowFunction
0..*
1..*
ParameterToWindowFunction
+windowFunctionParameter
+parameterValue
0..*
SpectralQuantitationParameters
+spectralQuantitationType
+spectralQuantitationAlgorithm
+spectralQuantitationParameters
+manualSpectralQuantitation

2.7. Data Sets

The content of the datasets described here is based on the JCAMP-DX format for NMR.

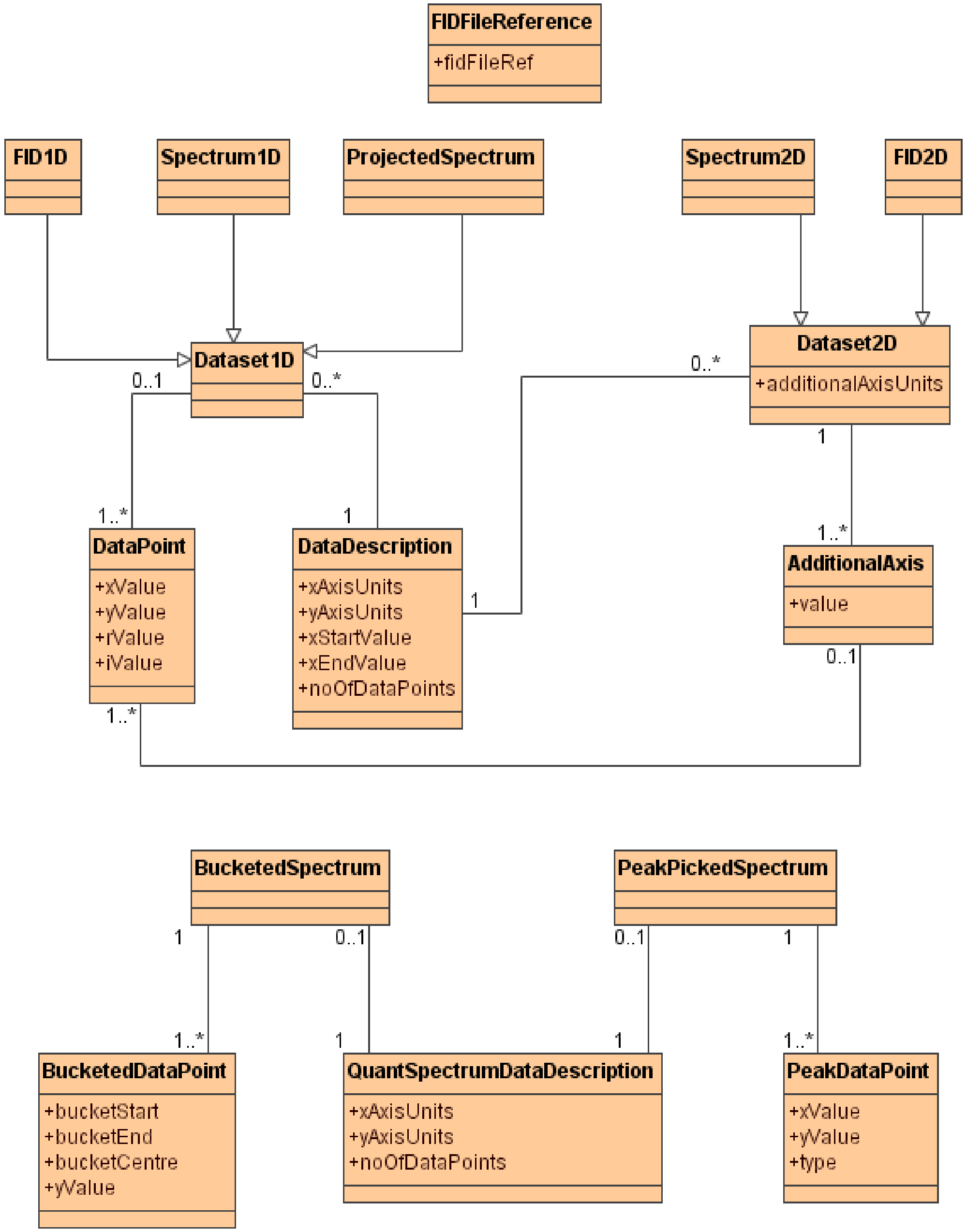
Block	Field				
	Name	**Definition**	**Domain**	**Units**	**?**
1D FID Data Set					
	xAxisUnits	The units of measurement on the x-axis	string	('second' | 'millisecond' | 'microsecond')	R
	yAxisUnits	The type of y-axis values	See §4.13		R
	xStartValue	The starting x-axis value.	float		R
	xEndValue	The final x-axis value.	float		R
	numberofDataPoints	The number of data points on the x-axis.	integer		R
	data matrix	The data matrix for the FID represented as either a set of y-axis values at equal x-axis intervals or a set of (x,y) pairs			R
FID File Reference					
	fidFileRef	A reference to a file that contains FID data.	URI		R
2D FID Data Set					
	additionalAxisUnits	The units of measurement on the second dimension axis.	string	('second' | 'millisecond' | 'microsecond' | 'hertz' | 'kilohertz' | 'megahertz')	R
	xAxisUnits	The units of measurement on the first dimension x-axis	string	('second' | 'millisecond' | 'microsecond')	R
	yAxisUnits	The type of values on the first dimension y-axis	See §4.13		R
	xStartValue	The starting value for the first dimension x-axis	float		R
	xEndValue	The final value on the first dimension x-axis	float		R
	numberofDataPoints	The number of data points on the x-axis in the first dimension.	integer		R
	data matrix	The data matrix for the FID which comprises multiple 1D FIDs each annotated with a value for the second dimension axis. Each 1D FID is represented as either a set of y-axis values at equal x-axis intervals or a set of (x,y) pairs			R
1D Spectrum					
	xAxisUnits	The units of measurement on the x-axis	string	('hertz' | 'kilohertz' | 'megahertz')	R
	yAxisUnits	The type of y-axis values	See §4.13		R
	xStartValue	The starting x-axis value.	float		R

<table>
<tr><th>Block</th><th colspan="5">Field</th></tr>
<tr><th></th><th>Name</th><th>Definition</th><th>Domain</th><th>Units</th><th>?</th></tr>
<tr><td></td><td>xEndValue</td><td>The final x-axis value.</td><td>float</td><td></td><td>R</td></tr>
<tr><td></td><td>numberofDataPoints</td><td>The number of data points on the x-axis.</td><td>integer</td><td></td><td>R</td></tr>
<tr><td></td><td>data matrix</td><td>The data matrix for the spectrum represented as either a set of y-axis values at equal x-axis intervals or a set of (x,y) pairs</td><td></td><td></td><td>R</td></tr>
<tr><td colspan="6">2D Spectrum</td></tr>
<tr><td></td><td>additionalAxisUnits</td><td>The units of measurement on the second dimension axis.</td><td>string</td><td>('hertz' | 'kilohertz' | 'megahertz')</td><td>R</td></tr>
<tr><td></td><td>xAxisUnits</td><td>The units of measurement on the first dimension x-axis</td><td>string</td><td>('hertz' | 'kilohertz' | 'megahertz')</td><td>R</td></tr>
<tr><td></td><td>yAxisUnits</td><td>The type of values on the first dimension y-axis</td><td>See §4.13</td><td></td><td>R</td></tr>
<tr><td></td><td>xStartValue</td><td>The starting value for the first dimension x-axis</td><td>float</td><td></td><td>R</td></tr>
<tr><td></td><td>xEndValue</td><td>The final value on the first dimension x-axis</td><td>float</td><td></td><td>R</td></tr>
<tr><td></td><td>numberofDataPoints</td><td>The number of data points on the x-axis in the first dimension.</td><td>integer</td><td></td><td>R</td></tr>
<tr><td></td><td>data matrix</td><td>The data matrix for the spectrum which comprises multiple 1D spectra each annotated with a value for the second dimension axis. Each 1D spectrum is represented as either a set of y-axis values at equal x-axis intervals or a set of (x,y) pairs</td><td></td><td></td><td>R</td></tr>
<tr><td colspan="6">2D Projected Spectrum</td></tr>
<tr><td></td><td>xAxisUnits</td><td>The units of measurement on the x-axis</td><td>string</td><td>('hertz' | 'kilohertz' | 'megahertz')</td><td>R</td></tr>
<tr><td></td><td>yAxisUnits</td><td>The type of y-axis values</td><td>See §4.13</td><td></td><td>R</td></tr>
<tr><td></td><td>xStartValue</td><td>The starting x-axis value.</td><td>float</td><td></td><td>R</td></tr>
<tr><td></td><td>xEndValue</td><td>The final x-axis value.</td><td>float</td><td></td><td>R</td></tr>
<tr><td></td><td>numberofDataPoints</td><td>The number of data points on the x-axis.</td><td>integer</td><td></td><td>R</td></tr>
<tr><td></td><td>data matrix</td><td>The data matrix for the spectrum represented as either a set of y-axis values at equal x-axis intervals or a set of (x,y) pairs</td><td></td><td></td><td>R</td></tr>
<tr><td colspan="6">Bucketed Spectrum</td></tr>
<tr><td></td><td>xAxisUnits</td><td>The units of measurement on the x-axis</td><td>string</td><td>('hertz' | 'kilohertz' | 'megahertz' | 'ppm')</td><td>R</td></tr>
<tr><td></td><td>yAxisUnits</td><td>The type of y-axis values</td><td>See §4.13</td><td></td><td>R</td></tr>
<tr><td></td><td>numberofDataPoints</td><td>The number of buckets in the spectrum.</td><td>integer</td><td></td><td>R</td></tr>
<tr><td></td><td>data matrix</td><td>The data matrix for the spectrum. The points in the data matrix comprise (the</td><td></td><td></td><td>R</td></tr>
</table>

Block	**Field**				
	Name	**Definition**	**Domain**	**Units**	**?**
		starting x-axis value for the bucket, the ending x-axis value for the bucket, the x-axis value at the centre of the bucket, the y-axis value)			
Peak-picked Spectrum					
	xAxisUnits	The units of measurement on the x-axis	string	('hertz' | 'kilohertz' | 'megahertz' | 'ppm')	R
	yAxisUnits	The type of y-axis values	See §4.13		R
	numberofDataPoints	The number of peaks in the spectrum.	integer		R
	data matrix	The data matrix for the spectrum. The points in the data matrix comprise (the x-axis value for the peak, the y-axis value for the peak, the type of the peak). For values for the type of a peak see §4.14			R

Association rules. There are no associations between the data blocks for the Data Sets concept.

Figure 7. UML for the Data Sets concept


FIDFileReference
+fidFileRef
FID1D
Spectrum1D
ProjectedSpectrum
Spectrum2D
FID2D
Dataset1D
0..1
0..*
Dataset2D
+additionalAxisUnits
0..*
1
1..*
DataPoint
+xValue
+yValue
+rValue
+iValue
1..*
1
DataDescription
+xAxisUnits
+yAxisUnits
+xStartValue
+xEndValue
+noOfDataPoints
1
1..*
AdditionalAxis
+value
0..1
BucketedSpectrum
1
0..1
PeakPickedSpectrum
0..1
1
1..*
BucketedDataPoint
+bucketStart
+bucketEnd
+bucketCentre
+yValue
1
1
QuantSpectrumDataDescription
+xAxisUnits
+yAxisUnits
+noOfDataPoints
1..*
PeakDataPoint
+xValue
+yValue
+type

3. Generating a Complete Description of an NMR Analysis

Using the data items in the tables above descriptions of various aspects of an NMR analysis can be created. To generate a complete description of an NMR analysis that is compliant with the reporting requirements, descriptions of the various concepts, in the terms described above, must be combined as described below:

- An Analysis must be associated with:
  - One NMR Sample
  - One Instrument
  - One Acquisition Parameter Set (of appropriate dimensions)
  - One Quality Control description
  - One or more of:
    - One FID Data Set which may be either:
      - A 1D FID Data Set
      - A 2D FID Data Set
      - A FID File Reference
    - Zero or many spectra which may be either:
      - A 1D Spectrum which is associated with:
        - A FID & Spectral Processing Parameter Set (for 1D NMR)
      - A 2D Spectrum which is associated with:
        - A FID & Spectral Processing Parameter Set (for multi-dimensional NMR without spectral projection). Note that if the Acquisition Parameter Set *pulseSequence* data item contains the value *2D J-Resolved* then the FID & Spectral Processing Parameter Set must be associated with a set of 2D J-Resolved Processing Parameters.
      - A 2D Projected Spectrum which is associated with:
        - A FID & Spectral Processing Parameter Set (for multi-dimensional NMR with spectral projection). Note that if the Acquisition Parameter Set *pulseSequence* data item contains the value *2D J-Resolved* then the FID & Spectral Processing Parameter Set must be associated with a set of 2D J-Resolved Processing Parameters.
    - Zero or many spectra resulting from spectral quantitation which may be either:
      - A Bucketed Spectrum which is associated with:
        - A FID & Spectral Processing Parameter Set
        - A Spectral Quantitation Parameter Set in which the *spectralQuantitationType* data item contains the value *bucketing*
      - A Peak-Picked Spectrum which is associated with:
        - A FID & Spectral Processing Parameter Set
        - A Spectral Quantitation Parameter Set in which the *spectralQuantitationType* data item contains the value *peak-picking*

4. Controlled Vocabularies

Some data items in the tables above are restricted to values from controlled vocabularies. These vocabularies are currently defined as follows. Extension of these vocabularies would represent a development of the standard.

4.1. Concentration standard type

- internal
- external

4.2. Sample introduction methods

- tube
- MAS
- flow probe

4.3. Water suppression

- Presat
- NOESY-Presat
- Watergate
- WET
- excitation sculpting

4.4. Pulse sequence name

- 1D
- 1D CPMG
- 2D J-resolved
- 2D TOCSY
- 2D Hadamard TOCSY
- 1D Diffusion Edited

4.5. Encoding

- TPPI
- States
- States-TPPI
- Quadrature filter
- Hadamard
- Radon
- GFT
- Frydman
- Echo/Anti-Echo

4.6. Post acquisition water suppression

- convolution
- polynomial fitting
- WaveWat
- HSVD

4.7. Transformation type

- fourier transformation
- non-fourier transformation

4.8. Window function

- exponential multiplication
- gaussian broadening
- sine
- sine^2

4.9. Window Function Parameter

- line broadening
- line sharpening
- sine bell length
- sine bell shift

4.10. Method of spectral projection

- maximum intensity
- summation

4.11. Spectral projection axis

- f1
- f2

4.12. Spectral Quantitation type

- peak picking
- bucketing

4.13. Y-axis value type

- power
- magnitude
- real
- imaginary
- complex

4.14. Peak-picked data point type

- singlet

- doublet
- triplet
- quadruplet
- multiplet
- unassigned