Conversion from metabolomics raw data to open format: ensuring MS and MS/MS data quality and software compatibility

Quentin Ruin, Delphine Centeno, Charlotte Joly, Marie Lagree, Stéphanie Durand, Emilien Jamin, Carla Orlandi, François Fenaille, Mélanie Pétéra, Estelle Pujos-Guillot

Conversion from metabolomics raw data to open format: ensuring MS and MS/MS data quality and software compatibility

Quentin Ruin[a], Delphine Centeno[a], Charlotte Joly[a], Marie Lagree[a], Stéphanie Durand[a], Emilien Jamin[b], Carla Orlandi[b], François Fenaille[c], Mélanie Pétéra[a], Estelle Pujos-Guillot[a]

[a] *Université Clermont Auvergne, INRAE, UNH, Plateforme d'Exploration du Métabolisme, MetaboHUB Clermont, Clermont-Ferrand, France*
[b] *Toxalim (Research Centre in Food Toxicology), Université de Toulouse, INRAE, ENVT, INP-Purpan, UPS, MetaboHUB, 31300 Toulouse, France*
[c] *Université Paris-Saclay, CEA, INRAE, Département Médicaments et Technologies pour la Santé (DMTS), MetaboHUB, 91191 Gif-sur-Yvette, France*

[1] The Galaxy Community. "The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update", Nucleic Acids Research, Volume 50, Issue W1, 5 July 2022, Pages W345–W351.
[2] Bruker Daltonics, Wissembourg, France
[3] DOI: 10.18129/B9.bioc.xcms [4] DOI : 10.1016/j.biocel.2017.07.002 [5] DOI: 10.1038/nbt.2377 [6] Avtonomov D.M. et al: J. Proteome Res. June 16, 2016. DOI: 10.1021/acs.jproteome.6b00021.
[7] DOI: 10.1093/bioinformatics/btq054 [8] https://doi.org/10.1038/s41587-020-0531-2

CONTEXT

In the context of information extraction of **high-throughput MS/MS metabolomics experiments, open science** has led to the necessity of **converting raw MS data into open formats** capable of handling MS/MS. However, several formats and conversion software exist, involving heterogeneous FAIR adherence in terms of **reproducibility, retro-compatibility and interoperability.**

MATERIAL AND METHODS

STUDIED FORMATS

Two major formats have become the standard for converting raw data: **mzML and mzXML**. Although netCDF is still notably in use, it can only store one specific MS-level as it is not originally designed for spectrometry and tends not to be supported by the latest software solutions.

Moreover, **netCDF is a mostly encoded format** with few to no metadata whereas **mzXML offers headers** containing acquisition and processing information (**Figure 1**). **MzML file metadata are even more detailed and structured,** in addition to information on each scan; only mass and intensity lists are encoded to reduce file size.

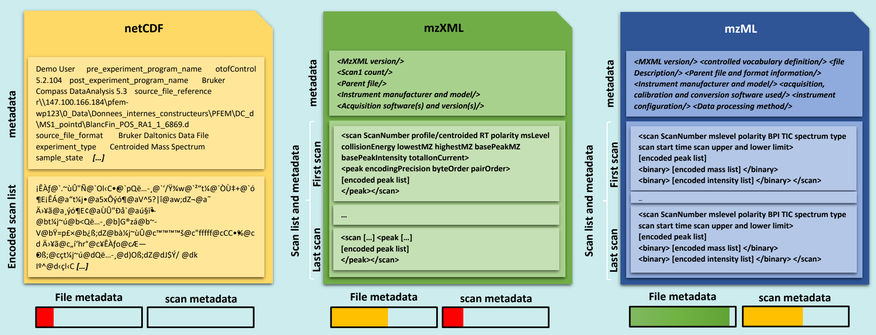


Figure 1: *Iconography of files when opened with a text editor. Normal text represents real raw content copied from a file as an example, bold text summarizes bigger content blocks or XML-like tags to enhance readability. The amount of human-readable metadata provided is shown underneath the file diagrams.*

In this poster, we focused on mzXML and mzML, as they are the most widely accepted by recent software solutions.

USED DATA AND SOFTWARE SOLUTIONS

Diverse data sources are essential to address format, manufacturer, instrument, and acquisition time issues. To evaluate **temporal reproducibility**, **software-dependent variations within the same format**, and the **compatibility of converted files with Galaxy**[1] and recent software solutions, we used data from a Bruker[2] Impact HDII UHR-QTOF. Analysis workflows in which the aforementioned datasets and software solutions are used, as well as their objectives, are outlined in the figure below.

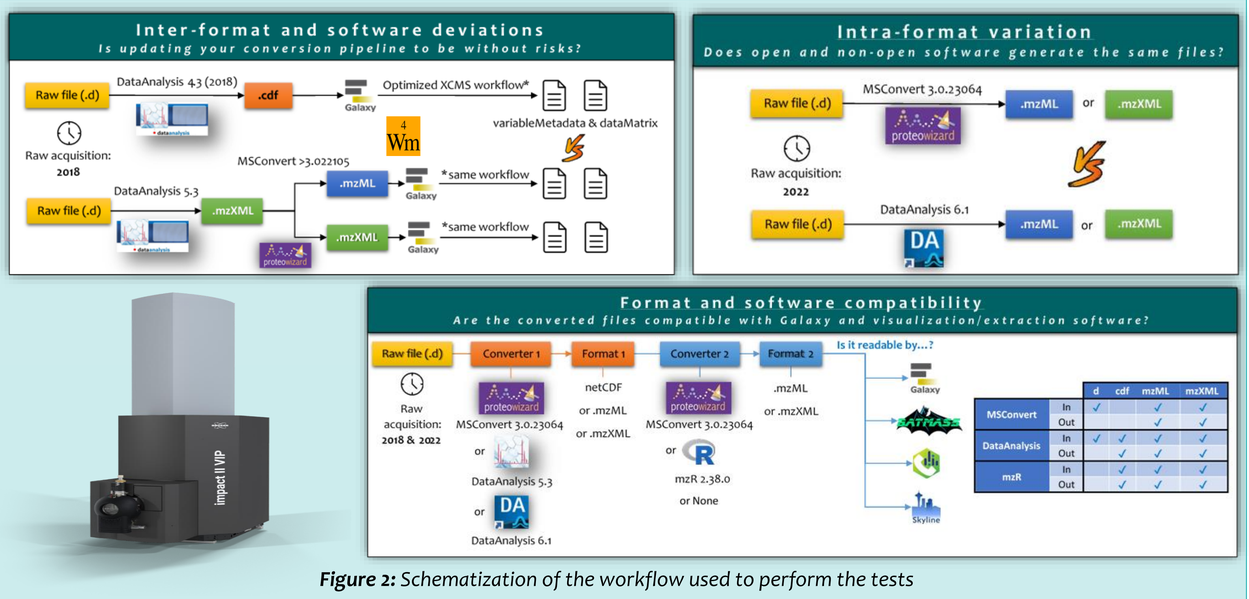


Figure 2: *Schematization of the workflow used to perform the tests*

DEDICATED TOOLS

To **streamline and automate** the necessary studies, we developed **two specialized Python tools:** ***mz(x)ml_compare*** **along with** ***XCMS_compare*** **(Figure 3).**

The **mz(x)ml_compare** tool is designed to **extract metadata from the headers of mzML and mzXML files,** compiling this information into a .tsv file without the need to open these large files in text editors. The **XCMS_compare tool analyzes the variableMetadata and dataMatrix** outputs from two XCMS[3] Galaxy workflows[4], **matching ions** by closest mass and retention time (RT) values (within a given window) to identify those detected in both workflows and providing **quality indicators that highlight potential deviations in mass, RT, or intensity values.**

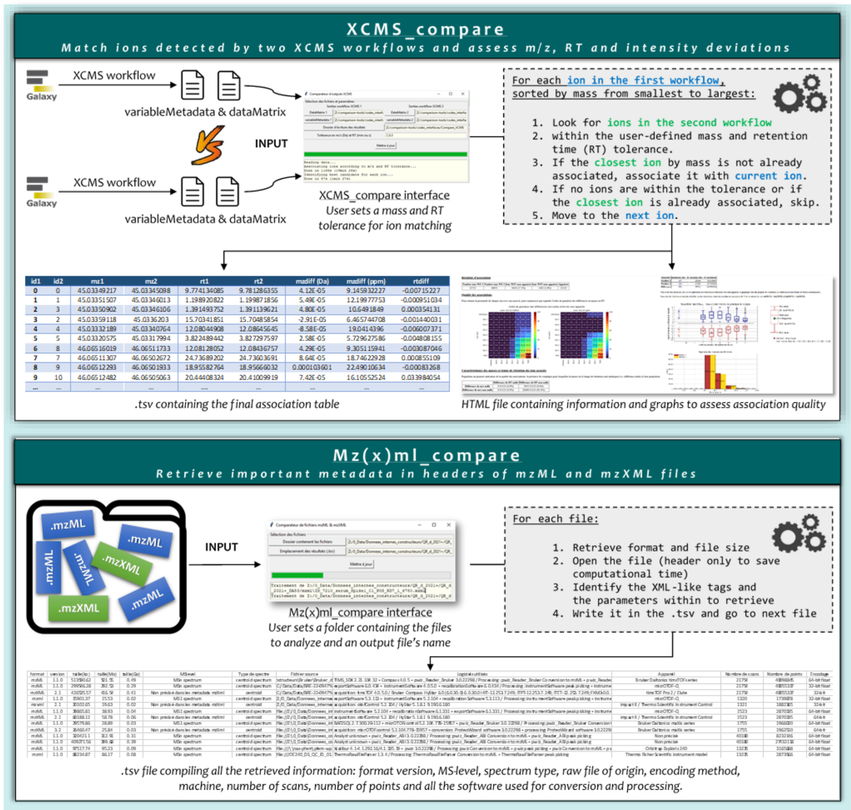


Figure 3: *Schematization of the workflow used to perform the tests*

RESULTS

INTRA-FORMAT VARIATIONS

Figure 4: *Iconography of files when opened with a text editor.*

Different ways to store acquisition software
- Different tag codes, content and organization

Different ways to store instrument reference
⚠ Instrument name is even false in both cases!

Different ways to store instrument components:
- Different number of components
- Different tag codes and organization
- Different "name" designation

Different scan metadata organization

Different encoding methods (even with the same encoding scheme, e.g. 64-bit)

Same format, different software: several metadata organization and potentially distinct encoding.

FORMAT AND SOFTWARE COMPATIBILITY

Knowing these variations within the same format depending on the conversion software used, especially with regard to **encoding**, **can all conversion methods produce files readable by Galaxy** and other common software for MS and MS/MS extraction/visualization? Tested with Galaxy, Batmass[6], Skyline[7] and MSDial[8] as target software (**Figure 5**).

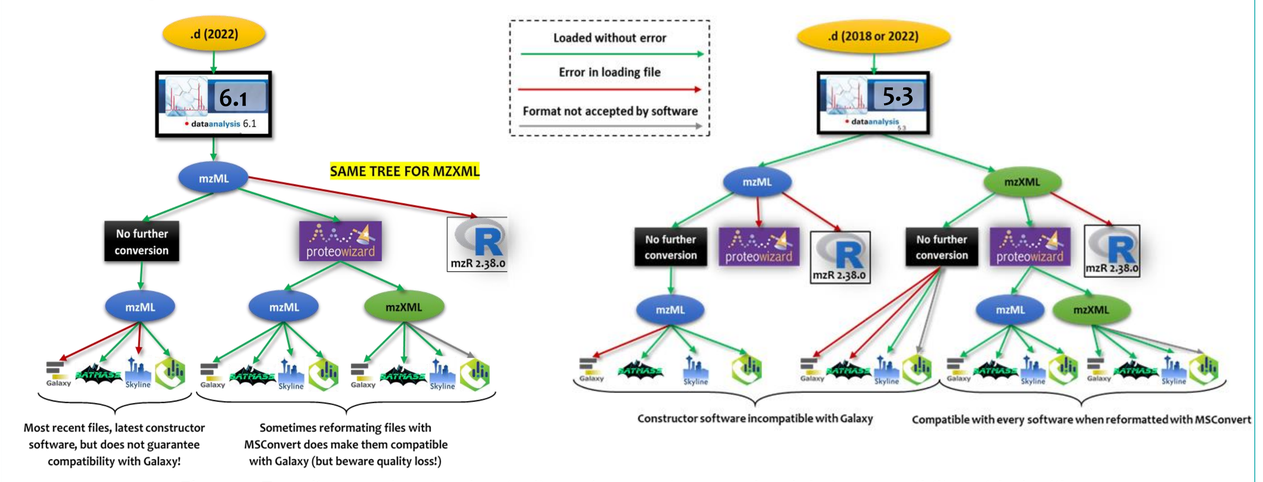
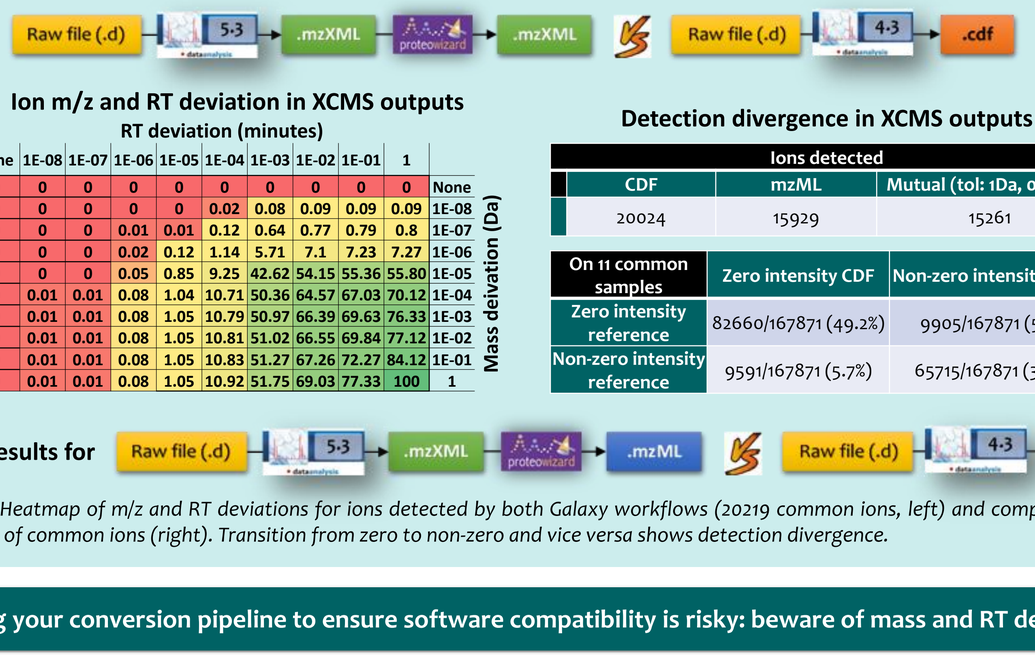


Figure 5: *Tree diagram showing the pipelines of conversion tested and their compatibility with the files.*

The compatibility of converted files with the used software must be monitored for all conversion software and each time a new version is released!

INTER-FORMAT AND SOFTWARE DEVIATIONS

As **DataAnalysis 5.3 did not ensure Galaxy compatibility** neither with mzML nor with mzXML (see **Figure 5** right tree), **we reformated the mzXML files with MSConvert** as the outputs are readable by Galaxy. However, for both mzML and mzXML outputs, **important detection divergences as well as mass and RT deviations can be observed** in comparison to the original netCDF file. In fact, **the number of detected ion differs** (20024 vs 15969) and **no mutually detected ions (15621) show identical mass and RT (Figure 6).** Moreover, **22.88% of them are deviated of at least $10^{-2}Da$ and 1 minute**, which are our maximal database matching tolerance. Worse, 11.6% show detection divergence, being detected in one workflow and not in the other.



None	1E-08	1E-07	1E-06	1E-05	1E-04	1E-03	1E-02	1E-01	1	
0	0	0	0	0	0	0	0	0	0	None
0	0	0	0	0	0.02	0.08	0.09	0.09	0.09	1E-08
0	0	0	0.01	0.01	0.12	0.64	0.77	0.79	0.8	1E-07
0	0	0	0.02	0.12	1.14	5.71	7.1	7.23	7.27	1E-06
0	0	0	0.05	0.85	9.25	42.62	54.15	55.36	55.80	1E-05
0	0.01	0.01	0.08	1.04	10.71	50.36	64.57	67.03	70.12	1E-04
0	0.01	0.01	0.08	1.05	10.79	50.97	66.39	69.63	76.33	1E-03
0	0.01	0.01	0.08	1.05	10.81	51.02	66.55	69.84	77.12	1E-02
0	0.01	0.01	0.08	1.05	10.83	51.27	67.26	72.27	84.12	1E-01
0	0.01	0.01	0.08	1.05	10.92	51.75	69.03	77.33	100	1

Figure 6: *Heatmap of m/z and RT deviations for ions detected by both Galaxy workflows (20219 common ions, left) and comparison of detection of common ions (right). Transition from zero to non-zero and vice versa shows detection divergence.*

Changing your conversion pipeline to ensure software compatibility is risky: beware of mass and RT deviations!

CONCLUSION AND PERSPECTIVES

Two popular open formats have become the norm to convert raw spectrometry data: mzML and mzXML. However, there are **several software solutions or software pipelines to obtain them** but metadata and, even more seriously, **masses, retention times and intensities can be subjected to non negligible deviations**. Moreover, there is **no certainty that converted files could be read by Galaxy** and other software until tests have been conducted. This is mainly due to the fact that **all software solutions do not write or read the same data encoding pattern**. This study revealed a valuable **imperative to ensure reproducibility: formats, dates, software and versions used must be monitored and reported!**

Now that we know how to obtain qualitative open format data, **the perspective of this study will focus on testing and selecting pipelines able to properly extract information from MS/MS** high throughput data.

Acknowledgments: This work is supported by the French Ministry of Research and National Research Agency as part of the French metabolomics and fluxomics infrastructure (MetaboHUB: ANR-11-INBS-0010).

quentin.ruin@inrae.fr

INRAE Centre Auvergne Rhône Alpes
Site de Theix
63122 Saint Genès Champanelle