

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

▶ To cite this version:

Quentin Ruin, Delphine Centeno, Charlotte Joly, Marie Lagree, Stéphanie Durand, et al.. Conversion from metabolomics raw data to open format: ensuring MS and MS/MS data quality and software compatibility. 16èmes Journées Scientifiques du Réseau Francophone de Métabolomique et de Flux-omique, Jun 2024, St Malo, France. hal-04611713

HAL Id: hal-04611713 https://hal.inrae.fr/hal-04611713

Submitted on 13 Jun2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



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.



INTRA-FORMAT VARIATIONS

mzML from DataAnalysis 6.1	mzML from MSConvert 3.0.23064	Different ways to store acquisition software							
METADATA	METADATA	 Different tag codes content and 							
<software id="instrumentSoftware" version="5.2.104"> <cvparam accession="MS:1000692" cvref="MS" name="Bruker software"></cvparam> <userparam name="AcquisitionProgram" value="otofControl"></userparam> </software>	<software id="micrOTOFcontrol" version="5.2.104.778-15957"> <cvparam accession="MS:1000726" cvref="MS" name="micrOTOFcontrol" value=""></cvparam> </software>	organization							
<cvparam accession="MS:1000703" cvref="MS" name="micrOTOF-Q"></cvparam>	<pre></pre>								
<componentlist count="3"></componentlist>	<componentlist count="5"> <source order="1"/> <cvparam accession="MS:1000073" cvref="MS" name="electrospray ionization" value=""></cvparam> <cvparam accession="MS:1000057" cvref="MS" name="electrospray inlet" value=""></cvparam></componentlist>	A Instrument name is even false in both cases!							



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

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 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 file information into a .tsv 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 values (within a given (RT) to identify those window) detected in both workflows and providing quality indicators that highlight potential deviations in mass, RT, or intensity values.



Most recent files, latest constructor Sometimes reformating files with MSConvert does make them compatible software, but does not guarantee with Galaxy (but beware quality loss!) compatibility with Galaxy!

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.



n m/z and RT deviation in XCMS outputs											
RT deviation (minutes)											
		1	1E-01	1E-02	1E-03	1E-04	1E-05	1E-06	1E-07	E-08	
	None	0	0	0	0	0	0	0	0	0	
	1E-08	0.09	0.09	0.09	0.08	0.02	0	0	0	0	
	1E-07	0.8	0.79	0.77	0.64	0.12	0.01	0.01	0	0	
5	1E-06	7.27	7.23	7.1	5.71	1.14	0.12	0.02	0	0	
q	1E-05	55.80	55.36	54.15	42.62	9.25	0.85	0.05	0	0	
2	1E-04	70.12	67.03	64.57	50.36	10.71	1.04	0.08	0.01	0.01	
ă.	1E-03	76.33	69.63	66.39	50.97	10.79	1.05	0.08	0.01).01	
2	1E-02	77.12	69.84	66.55	51.02	10.81	1.05	0.08	0.01).01	
	1E-01	84.12	72.27	67.26	51.27	10.83	1.05	0.08	0.01	0.01	
_	1	100	77.33	69.03	51.75	10.92	1.05	0.08	0.01	0.01	

ection divergence in XCMS outputs

lons detected								
CDF	mzML	Mutual (tol: 1Da, 0.3 min)						
20024	15929	15261						
On 11 common samples	Zero intensity CDF	Non-zero intensity mzML						
Zero intensity reference	82660/167871 (49.2%)	9905/167871 (5.9%)						
Non-zero intensity reference	9591/167871 (5.7%)	65715/167871 (39.1%)						

- wzwr	mzXM	·m; /IL	.m	EML	U t	Mz(x)ml_ o analyze and	compare interfai lder containing th an output file's	ce name	2. 0 3. 1 4. k	Open th Computa Cdentif Che par Irite i	e file (header on tional time) y the XML-like ta ameters within to t in the .tsv and	ly to s gs and retrie go to	ave ve next fi	le
ormat samin	n telleller	1 talle(Ma)	bille(Cr)	Manal	Tune de marite	Diffierman		Logidal utiliste			Armoil	No others de s cares	Northos de cointr	Forstum
maML 11.0	511990.0	2 301.35	0.49	MSc spectrum	centrol dispectrum	http://www.inview.com	TMS_SDK 2_21_104_32 + Comparts 4.0, 5 + out	Bander Brukar 30,22290 / Proceed	ne : awit Reader, Bruker Comercian to rad	ML+ quit Reader	Bruier Daitonics timeTOEs eries	2175	40090045	G1-bit float
maML 11.0	2999162	9 292.50	0.29	MSn spectrum	centrol dispectrum	C/Data/Data/08E-22-0947% et	e portSoftware 6.0. (3) + instrumentSoftware	4.0.5.0 + recalibrationSofte are 6.0.4	34 / Processing : instrumentSoftware peak o	aidéne + instrumei	micrOTOF-0	2175	405510	32-bit float
m:004E 2.1	426205.5	7 416.51	0.41	Non coèdré dans les metadats miXmi	pentroid	C/Data/Data/080-2240947% at	a quisition: trm TCE 4.0.5.0 / Bruler: Compar	11/5br 60(6030.06.0300)/IT-12	253 7 249 ITEE 12 253 7 249 ITEE 12 253	7249-53863-0.0	Hime TOE Pro 2 / Elute	2175	4055317	12-b t
maml 1.1.0	2901.3	7 1553	0.02	NS1 spectrum	centrol dispectrum	Z/O Data/Donnees Internet et	exportSoftware 5.3.10 + instrumentSoftware	5.2.104 + recallb ration Software 5.3.1	13 / Processing instrumentSoftware peaks	o icking + instrume	micrOTOF-0	1320	1739678	32-bit float
museril 2.1	20102.6	5 19.63	0.02	Non prédité dans les metadats mXmi	awntroid	Z/O Data/Donnees, internet a	a quilitar: ab Control 52, 104 / HySter 5, 17	1 9 190 6 180			imp at II / Therma Scientific in strument Control	1121	1887 15	12-bit
maML 11.0	30005.0	1 30.93	0.04	MS1 spectrum	centrol dispectrum	Ne: 1/7/10 Date/Donnees int in	Instrument Software 5,2104 + recallbratio rSt	ftware 6.1.331 + exportSoftware 6.1.3	11 / Processi or: InstrumentSoftware peaks	o ickine + instrume	micrOTDE-0	1523	2070305	G1-bit float
m:XIVE 2.1	@188.1	2 58.78	0.06	Non prédité dans les metadats mXml	avritraid	Hec//0/0 Date/Donnees, inf a	caultitor: ab Control 52, 101 / HySter 5, 1/	1 9 190 6 190			implact II / Thermo Scientific in strument Control	1523	2870.05	Gl-bit
maML 11.0	2679.0	6 28.89	0.03	NS1 spectrum	centrol dispectrum	He://7.10 Dab/Donneet, int B	M250L2.730020112 + micrOTOR antral 5	2.108.728-15957+pwiz Reader Brui	ker 10.2290 / Processing: gwit Render D	ruker Conversion (Bruker Dalitani ca makia series	1755	196620	GI-bit float
m20ML 3.2	25460.4	7 25.84	0.03	Non prédaé dansies metadata mXmi	oentroid	He://7.10 Date/Donnees. Inf a	a guilti tan: micrOTOFcontrol 52.104.779-E/	57 + convention: ProteoWizard softwa	are 3.0.22298 + properting ProteoWit and a	offware 30,2229	Bruker Dalitonica, makia, series,	1755	1962/20	Glbit
maML 1.1.0	30421	1 312.91	0.31	MSn spectrum	centrol dispectrum	His://210 Dats/Donnest on A	Analyst unknown + pwiz Reader A013.0227	80 / Processing zwiz Reader ABI Con	wention to moML+ pwiz Reader ADIp eak	pic lên e	Nan ankcisk	401.00	823236	G4-bit float
maML 11.0	4092715	2 39.00	0.19	Min spectrum	centrol dispectrum	Flec//7:\0 Date/Dannees, as A	Analyst unknown + pwiz Reader A013.0227	80 / Processing grul z Reader Abi Con	wendon to moML + pwiz Reader Allipeak	picián g	Nan précisé	40130	2702110	GI-bit float
maML 1.1.0	9517.7	8 95.23	0.09	MSn spectrum	centrol dispectrum	file: //// \nas-ofern\ plem-wpl X	kalbur 4, 1-4, 1, 292 16/4, 1, 305, 39 + gwit 30	22290 / Processing zwit Conversion t	to maML + a with a welk aic lither + aw it Co merc	tion to moML+ a v	Orbit ap Basianis 240	13235	3165/28	G4-bit float
maml 1.1.0	90234.0	7 86.17	0.08	MSn spectrum	centrol dispectrum	He:///OE240 DS OC ID 01 7	hermoRay/RieParter 1.3.4 / Processing The	moRawRiePars er Conversion to moM	L + ThermpRawFileParser peak picking		Thermo Rober Scientific Instrument model	132B	2073266	G4-bit float
mami 11.0) 8 234.8	7 86.17	.tsv j	file compiling all t machine, r	he retriev number o	ved informat f scans, num	ion: format versions and interview to the second seco	on, MS-level, sp all the softwa	ectrum type, rav re used for conve	v file of rsion ar	origin, encoding met nd processing.	hod,	2073926	GI-bit float

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



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

None

0

Two popular open formats have become the norm to convert raw spectrometry data: mzML. However, there are several 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







