PSI Recommendation
Status: DRAFT

PSI Mass Spectrometry Standards Working Group

Eric W. Deutsch, Institute for Systems Biology
Juan Antonio Vizcaíno, EMBL-EBI
Yasset Perez-Riverol, EMBL-EBI
Jeremy Carver, University of California San Diego
Shin Kawano, Database Center for Life Science
Pierre-Alain Binz, Centre Hospitalier Universitaire Vaudois
Benjamin Pullman, University of California San Diego
Jim Shofstahl, Thermo Fisher Scientific
Ralf Gabriels, Ghent University
Tim Van Den Bossche, Ghent University
Nuno Bandeira, University of California San Diego

May 20, 2021

USI: A Universal Spectrum Identifier for Mass Spectrometry

Status of this document

This document provides information to the proteomics community about a universal spectrum identifier for proteomics and metabolomics and other fields that use mass spectra. Distribution is unlimited.

Version Draft 10 - this is a draft of version 1.0

Abstract

The Human Proteome Organization (HUPO) Proteomics Standards Initiative (PSI) defines community standards for data representation in proteomics to facilitate data comparison, exchange and verification. This document presents a specification for a multi-part identifier of the form mzspec:<collection>:<msRun>:<indexType>:<indexNumber>:<optional interpretation> for mass spectra so that they may be easily and universally referenced for subsequent access. The Universal Spectrum Identifier (USI) describes a virtual path to locate a spectrum plus a possible interpretation of that spectrum. The USI is being implemented at most ProteomeXchange partner repositories and can be freely used by any other software. While initially implemented for the field of proteomics, its design is amenable to other fields that use mass spectra for analyte detection. Further detailed information, including any updates to this document, implementations, and examples is available at http://www.psidev.info/usi.

Contents

1 Introduction

1.1 Description of the need

There is currently a problem in the field of proteomics when mass spectra are used as evidence for an important scientific finding. Many journal guidelines require that such findings be supported by an annotated spectrum in an effort to demonstrate the correctness of the identification (1–3). This is a positive advance over not showing any evidence, but is often unsatisfying to reviewers and readers because the annotated spectrum depiction is often of modest quality, perhaps a pixelated screenshot embedded in a PDF. Furthermore, even when the depiction is of good quality, the static representation resists any efforts to explore alternate identifications or assess apparently unannotated peaks. Another common scenario is that a finding is described in one publication along with a supporting depiction, but then a follow-up reprocessing of the data set does not result in a corroborating identification. Without an identifier attached to the original claim, it is difficult to trace back to determine whether the original analysis or the new analysis is in error. Additionally, it would be beneficial that spectra without annotation also have a unique identifier by which they can be referred to. An efficient mechanism to refer to both spectra and their potential annotations is needed.

1.2 Requirements

The main requirements to be fulfilled for a universal spectrum identifier (USI) are:

- The identifier should be a multi-part key that encodes the abstract path needed to identify a specific spectrum contained in a specific collection of spectra.

- The identifier should be creatable without a special algorithm other than to concatenate fairly common attributes of a spectrum.

- Authors of research articles should be able to generate identifiers for individual spectra before they submit the article.

- The identifier should be unambiguously parsable and interpretable into its separate components.

- The identifier components should enable the receiver of an identifier to locate the spectrum without the requirement for pre-generating identifiers for all spectra contained in the resource.

- The identifier should be applicable to spectra that were generated both before and after the final ratification of the identifier standard.

- The identifier should be a text string that can be easily embedded in URLs or copy-pasted into web forms.

- The identifier should have some level of potential to be fixed by humans or algorithms if it is provided in a slightly garbled or incomplete way.

- The identifier should be primarily designed to meet the needs and customs of the proteomics community, but should in principle be applicable to other communities that use mass spectra for their analysis such as metabolomics.

- The identifier should not require a central managing resource to mint identifiers, although, to be most effective, it may benefit from coordinated minting of collection identifiers.

1.3 Issues to be addressed

The main issues to be addressed by the identifier are:

- There is no standard mechanism that can be interpreted and applied in the same way across multiple proteomics resources for providing a unique identifier for any mass spectrum, and potentially its interpretations.

2 Notational Conventions

The key words “MUST“, “MUST NOT”, “REQUIRED”, “SHALL”, “SHALL NOT”, “SHOULD”, “SHOULD NOT”, “RECOMMENDED”, “MAY”, and “OPTIONAL” are to be interpreted as described in RFC 2119 (4).

3 The Universal Spectrum Identifier (USI) Definition

3.1 The documentation

The documentation of the USI is divided into several components. All components in their most recent form are available at the main USI page on the HUPO-PSI website (http://www.psidev.info/usi).

- Main specification document (this document).

- List of current implementations with examples (section 3.3.7).

Throughout the term “resource” is used to refer to an automated system that can receive a USI, interpret it, and respond with information. The term “user” is used to refer to person or automated system that uses a USI to request information from a resource.

3.2 Relationship to other specifications

The specification described in this document is not being developed in isolation; indeed, it is designed to be complementary to, and thus used in conjunction with, several existing and emerging models. Related specifications include the following:

1 *PSI Spectral Library Format* (http://www.psidev.info/mzSpecLib) The PSI Spectrum Library Format is being developed as a standard mechanism for storing spectrum libraries. Many spectrum library entries are derived from multiple spectra, and this provenance will be referenced using USIs.

2 *PROXI* (http://www.psidev.info/proxi). The Proteomics Expression Interface being developed by the PSI is a standardized API by which mass spectrometry proteomics information can be exchanged. References to individual spectra will be made via USIs.

3 *mzML* (http://www.psidev.info/mzML). mzML is the PSI open data standard for MS data (5) and is one of the file formats that can be referenced in USIs.

4 *ProForma 2.0* (http://www.psidev.info/proforma). While the present specification focuses on an identifier for a spectrum, a companion specification, called ProForma 2.0, describes the notation for describing the potential interpretation of a spectrum.

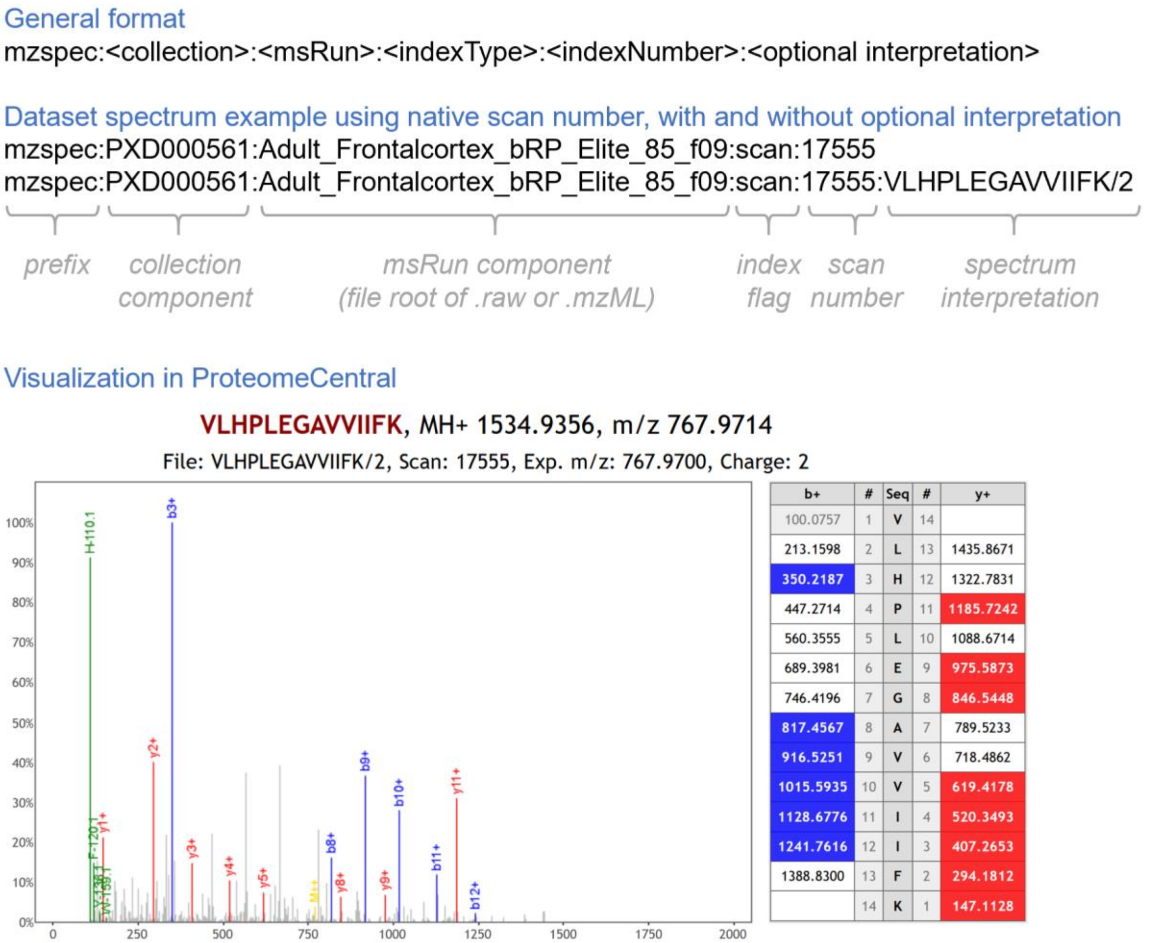
5 SPLASH (Spectral Hash Identifier) (https://splash.fiehnlab.ucdavis.edu/). A SPLASH is a mechanism for computing a hash for one instance of a spectrum so that identical versions of that same spectrum can be identified across different repositories. It is designed specifically for small molecule reference spectra.

6 mzTab (http://psidev.info/mzTab) The PSI mzTab specification for peptide and protein identifications and quantification has references to spectra and spectrum files. It does not currently support USIs, but perhaps future versions will if there is a need.

7 mzIdentML (http://psidev.info/mzIdentML) The PSI mzIdentML specification for peptide and protein identifications has references to spectra and spectrum files. It does not currently support USIs, but perhaps future versions will if there is a need.

3.3 The basic form of the USI

The Universal Spectrum Identifier (USI) is a multi-part key providing an abstract path for each spectrum, with specially defined fields with increasing levels of detail from left to right. The identifier has a basic form for a plain uninterpreted spectrum, and then several extensions for interpretations and other related concepts. First the components of the basic form are introduced and then the various related extensions and complications are described in later sections of this specification. The basic form is as depicted in Figure 1.



General format

mzspec:<collection>:<msRun>:<indexType>:<indexNumber>:<optional interpretation>

Dataset spectrum example using native scan number, with and without optional interpretation

mzspec:PXD000561:Adult_Frontalcortex_bRP_Elite_85_f09:scan:17555

mzspec:PXD000561:Adult_Frontalcortex_bRP_Elite_85_f09:scan:17555:VLHPLEGAVVIIFK/2

Visualization in ProteomeCentral

VLHPLEGAVVIIFK, MH+ 1534.9356, m/z 767.9714

File: VLHPLEGAVVIIFK/2, Scan: 17555, Exp. m/z: 767.9700, Charge: 2

b+	#	Seq	#	y+
100.0757	1	V	14	
213.1598	2	L	13	1435.8671
350.2187	3	H	12	1322.7831
447.2714	4	P	11	1185.7242
560.3555	5	L	10	1088.6714
689.3981	6	E	9	975.5873
746.4196	7	G	8	846.5448
817.4567	8	A	7	789.5233
916.5251	9	V	6	718.4862
1015.5935	10	V	5	619.4178
1128.6776	11	I	4	520.3493
1241.7616	12	I	3	407.2653
1388.8300	13	F	2	294.1812
	14	K	1	147.1128

Figure 1: Presentation of the general format of the Universal Spectrum Identifier, including the mzspec prefix, the collection component, the msRun component, the index type, the index number, and the optional spectrum interpretation (top). A dataset spectrum example (mzspec:PXD000561:Adult_Frontalcortex_bRP_Elite_85_f09:scan:17555:VLHPLEGAVVIIFK/2) using the native scan number, once with and without the optional spectrum interpretation (middle). Visual representation of this spectrum at the ProteomeCentral USI search and validation interface (http://proteomecentral.proteomexchange.org/usi/) (bottom left), accompanied with the ion table indicating the m/z of the identified b- and y-ions (bottom right).

3.3.1 Component separators

USIs are multi-part keys with several components of increasing granularity from left to right. Each component is separated by the colon (:) character.

3.3.2 **The mzspec prefix**

The first component is a prefix of “mzspec”. This functions as a signal that is immediately recognizable to both humans and computer code to indicate that a string is a USI. The mzspec prefix has been registered with identifiers.org, a registry of such namespaces that enable the translation of CURIEs (6) to fully qualified URIs and URLs. The mzspec prefix MUST always be all lowercase.

3.3.3 The collection component

The second component is the data collection identifier. The goal is to support only the most universal identifiers, not individual repository identifiers, since these would not be universally understood by different resources. In general, PXD identifiers are preferred over RPXD identifiers, except in the case of the "PSM provenance identifier" (see below) where specification of the RPXD is obligatory when the provenance refers to a reprocessing result. In all other cases, the PXD is preferred since this will be more widely understood across repositories. Only if a PXD is not available for a particular dataset may a repository use a local identifier. There can be use cases when individual resource identifiers are appropriate, but this does detract from the "universal" part of the standard. A complete list of valid identifiers can be found on https://github.com/HUPO-PSI/usi/blob/master/CollectionIdentifiers.md.

In cases where a USI must be written at a time when the final collection identifier is not yet known, a placeholder of USI000000 SHOULD be used as the collection identifier, so as to enable a straightforward search and replace when the final PXD or other collection identifier is registered. Software implementations SHOULD recognize this string as a placeholder that needs to be resolved. For example, primary data analysis software might internally use a USI such as mzspec:USI000000:fraction24:scan:24922 to identify a particular scan in the "current dataset" (for which there is not a public identifier yet). At some later date when the dataset gains a public identifier, a global search and replace of USI000000 with the actual collection identifier can be performed. No other components can have a similar placeholder.

3.3.4 The msRun component

The third component is the mass spectrometry run name (msRun). This is most typically the root name of a .RAW file or .mzML file, and the folder name for Waters .raw/ folders and Agilent .d/ folders, etc. Although it was initially expected that certain characters would require some escaping, it was decided to disallow escaping of any characters. All characters present in the msRun name MUST be rendered as they are, without escaping. This could possibly cause some complications, but no disabling complications have been identified in any existing datasets. Even colon (:) characters in the msRun name will not be escaped since the following field (indexType; see below) is limited to only a few values ("scan", "index", "nativeId" or "trace"), and USI parsers are expected to detect and compensate for colons in the msRun name by triggering on the limited values in the next field <indexType> to detect embedded colons. This was decided because it seems easier and more reliable to write parsers that can handle this situation than getting users to reliably escape colons or other characters properly. Other fields will not have special characters, so only the msRun has this complication. In principle Unicode characters are allowed when necessary, but this has not been tested since examples in the wild were not found. This component is case sensitive, although resources receiving USIs MAY compensate for case mismatches as a courtesy to the user.

A basic assumption in this design is that all msRuns in a collection have unique names. As long as msRun names are unique within a collection, any possible path or subfolder MUST NOT be specified in the USI. It is understood that a resource resolving a USI may need to search around within the collection to find the msRun. Different resources sometimes organize datasets differently, so a folder path isn't likely to be universal, nor necessary. This is mostly true, but there are some datasets where we know that several different samples are organized in separate folders and the files are named fraction01.mzML, fraction02.mzML, etc. in each folder! If and only if this is the case with a certain collection, the msRun MUST be prefixed with "[subFolder]". See subFolder section below for more details. This is very rare.

It is encouraged that resources do not rename msRuns in their resource so that all msRuns can be referenced with a USI. However, if it is necessary that a resource renames a file (e.g. replacing spaces or other unusual characters with underscores to aid software that cannot handle spaces in filenames), the resource must maintain a lookup table to match original names and new names, so that original names can be used in all cases related to USIs.

The msRun name may be augmented with an extension such as .mzML, .RAW, .mgf or .d, which would reduce the universality of the USI, but may provide substantial benefits when communicating with an individual resource when multiple versions of an msRun data are available there. This becomes especially important when index numbers into peak list files that contain a subset of the original spectra, such as mgf, pkl, ms2 files, are provided. Resources supporting USIs MUST properly parse and understand such common extensions, but it is not stipulated in this specification how a resource must handle such extensions.

For SCIEX .wiff files, there is an added complication that multiple MS runs can be stored in a single file. During conversion to mzML, most converters create one mzML file per run, with the mzML file named after the embedded MS run name. Therefore, in such cases the USI MUST use the embedded MS run name rather than the .wiff file name.

3.3.5 The indexType component

The fourth component is the indexType, which specifies the type that follows in field five. Presently allowed values are: “scan”, “index”, “nativeId”, or “trace”. If at all possible, this SHOULD be “scan”, meaning the scan number in the original MS run as produced by the originating instrument. However, there are some potential complications for some vendors or in cases where pedigree has been lost due to e.g. conversion to MGF. For Thermo instruments, a value of “scan” means the “scan number”, not the “scan index”. Also, for the spectral library use case below, there are no scans, only index numbers.

For MS run files produced by a vendor where a single scan number does not uniquely identify a scan, the USI MUST use a compact nativeId mechanism to identify the specific scan. See section 3.5.4 for a complete description of this mechanism.

If only an MGF file is available and reference to the original scan is lost, or for spectral libraries where a “scan” is not relevant, then the indexType MUST be “index”. See section 3.5.2 for more information on this scenario.

The indexType field may also be “trace” to signify an SRM chromatogram within an msRun that contains SRM traces.

See below for further discussion of these values.

3.3.6 The indexNumber component

The fifth component is the indexNumber, which specifies the index number(s) of the spectrum in the previously defined entity (collection + msRun). If at all possible, this SHOULD be the scan number in the msRun. In cases where scan numbers are not available, another type of index may be substituted, such as for MGF files where the original scan information is lost, or when the indexType is “nativeId” (see section 3.5.4 for a discussion hereof).

3.4 Additional forms

This multi-part identifier described above is universally unique for all spectra made available in publicly deposited datasets and provides a framework for substantial extended functionality.

Although the overall name of this construct is the “Universal Spectrum Identifier”, there are several additional forms that come as a result of subtracting or adding to the identifier. These will be defined in the subsections below.

3.4.1 The MS Run Identifier

Implied within the USI is an MS run identifier. Every deposited MS run can be referenced with a shortened form:

mzspec:<collection>:<msRun>

such as this example:

mzspec:PXD000561:Adult_Frontalcortex_bRP_Elite_85_f09

Since an MS run may be represented in several formats (with potentially slightly different data associated with the spectrum), a format suffix MAY be specified:

mzspec:PXD000561:Adult_Frontalcortex_bRP_Elite_85_f09.RAW

to signify that the .RAW file is meant. Although each resource must recognize this extension for what it is, how a resource must handle this extension is not prescribed, as discussed above in subsection 3.3.4.

3.4.2 The PSM Identifier

Whereas the basic USI identifies only a spectrum, in many cases it will be useful to specify an interpretation of the spectrum (**which may or may not be correct**), i.e. a peptide-spectrum match (PSM). This form is:

mzspec:<collection>:<msRun>:<indexType>:<indexNumber>:<interpretation>

where <interpretation> will typically be <peptidoform>/<charge> (although can be other things in the case of cross-linking MS, metabolomics, lipidomics, etc.)

such as:

mzspec:PXD000561:Adult_Frontalcortex_bRP_Elite_85_f09:scan:17555:VLHPLEGAVVIIFK/2

This example corresponds to the spectrum from scan 17555 of the specified run that may be annotated to show evidence for doubly charged (signified by “/2”) peptide ion VLHPLEGAVVIIFK. The primary goal of the USI is to identify the spectrum/scan, but adding the interpretation suffix is potentially extremely powerful, and is therefore a feature that is supported from the beginning. Specifying mass modifications and other complex scenarios have been fully defined in a companion specification ProForma 2.0 (http://www.psidev.info/proforma).

The charge number MAY be negative for negative polarity scans. The charge MAY be 0 to signify an unknown charge, although it is not specified what a receiver of such an USI might do with that information. A possible use case for this is a very large window DIA scenario where fragment ions from the same peptidoform in different charge states may be detected and identifiable. Furthermore, the explicit requirement for a charge in the form of the “/n” suffix can help disambiguate possible USI truncation versus no charge information.

It is important to highlight that the PSM written into the <interpretation> field of the USI could be different from the one originally provided by the authors when the original dataset was submitted to one of the PX resources. The way this identifier is constructed in fact allows any peptide sequence to be represented, even clearly incorrect ones. When a given resource receives a PSM identifier, the precise behavior is not prescribed here. It would be ideal if the resource could alert the user on how the user-specified interpretation compares with information contained at the resource. For example, if the user provides an interpretation for a spectrum, it would be ideal for the resource to inform the user if local interpretation information agrees with or disagrees with the user-supplied interpretation. In many cases, this may be difficult and resources are not obligated to do this.

3.4.3 Specifying mass modifications in the peptide for encoding PSMs

Many peptide ions include a mass modification, making the identity of the interpretation more complicated than a string of amino acid letters. There are various other complications that could be entertained: multiple co-fragmenting peptides, chemical formulas, cross-linked spectra. All aspects of encoding an associated interpretation are fully described in the companion PSI ProForma (Proteoform and Peptidoform Notation) specification, a separate document. It is available at http://psidev.info/proforma.

3.5 Examples

In this section, we provide the reader with several URLs to existing USIs. Although these are functional at the time of this writing and although the USIs will not change, URLs could sometimes change over time in different resources. If these URLs no longer work at the time of reading, consult the main PSI USI web page for updates to the URLs: http://www.psidev.info/usi.

USI example 1
mzspec:PXD000561:Adult_Frontalcortex_bRP_Elite_85_f09:scan:17555:VLHPLEGAVVIIFK/2

PeptideAtlas	https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/ShowObservedSpectrum?USI=mzspec:PXD000561:Adult_Frontalcortex_bRP_Elite_85_f09:scan:17555:VLHPLEGAVVIIFK/2
MassIVE	https://massive.ucsd.edu/ProteoSAFe/usi.jsp#{"usi":"mzspec:MSV000079514:Adult_Frontalcortex_bRP_Elite_85_f09:scan:17555:VLHPLEGAVVIIFK/2"}

USI example 2
mzspec:PXD000966:CPTAC_CompRef_00_iTRAQ_12_5Feb12_Cougar_11-10-11.mzML:scan:11850:[UNIMOD:214]YYWGGLYSWDMSK[UNIMOD:214]/2

PRIDE	https://www.ebi.ac.uk/pride/archive/spectra?usi=mzspec:PXD000966:CPTAC_CompRef_00_iTRAQ_12_5Feb12_Cougar_11-10-11.mzML:scan:11850:[UNIMOD:214]YYWGGLYSWDMSK[UNIMOD:214]/2
PeptideAtlas	https://db.systemsbiology.net/sbeams/cgi/PeptideAtlas/ShowObservedSpectrum?usi=mzspec:PXD000966:CPTAC_CompRef_00_iTRAQ_12_5Feb12_Cougar_11-10-11.mzML:scan:11850:[UNIMOD:214]YYWGGLYSWDMSK[UNIMOD:214]/2
MassIVE	https://massive.ucsd.edu/ProteoSAFe/usi.jsp#{"usi":"mzspec:PXD000966:CPTAC_CompRef_00_iTRAQ_12_5Feb12_Cougar_11-10-11.mzML:scan:11850:[UNIMOD:214]YYWGGLYSWDMSK[UNIMOD:214]/2"}

USI example 3
mzspec:PXD005175:CRC_iTRAQ_06:scan:11803:VEYTLGEESEAPGQR/3

jPOST	https://repository.jpostdb.org/spectrum/?USI=mzspec:PXD005175:CRC_iTRAQ_06:scan:11803:VEYTLGEESEAPGQR/3

3.6 Complex Issues

The above specification provides a definition of the USI that will apply to most common cases. However, there are a set of known issues that complicate the specification. These complications are part of this formal specification, but the design decisions are not well tested and may need to be refined in future versions of this specification after the community has gained more experience with the common cases.

3.6.1 [subFolder] prefix for msRun

The [subFolder] construction is required to disambiguate instances where the same collection has two or more msRuns with the same name. This is rare, but does happen. For example, imagine PX dataset PXD123456 where there are two subfolders/files: CaCo01/A01_100ng.RAW and then also Ctrl01/A01_100ng.RAW, i.e. there are two MS runs from different samples with the same name in one PXD in different folders. Note that individual resources are responsible for finding the MS runs within their directory structure without hints from this field; this is easy as long as MS run names are uniquely named within a collection. Multiple levels of subfolders MUST be separated with forward slash ('/') characters. There MUST NOT be multiple bracketed subfolders (i.e. [subFolder1][subfolder2] is not permitted). The presence of a subfolder is denoted by the first character of the msRun being the open square bracket ('['). An open square bracket that is not the first character of the msRun is considered part of the msRun name. For an example of a single dataset containing different MS runs with the same filename in different folders, find the QC_2.mzML files at ftp://massive.ucsd.edu/MSV000085129/peak/. For SCIEX WIFF files it is possible to put multiple MS runs within one WIFF file, and the names of these runs need not match the WIFF file name roof. As long as all MS run names within a dataset is unique, there is no problem. However, it is possible that there are two different SCIEX wiff files within one dataset that contain MS runs within them. In this case, the [subFolder] mechanism can be used to provide the WIFF file name to disambiguate the situation. At the moment of writing, no examples of this are known, but this situation is possible.

If a subfolder contains a colon character (not generally allowed in Windows, but permissible under Linux) or square bracket characters, they are not escaped, and the onus is on all parsers to parse the subfolder correctly, allowing any characters between the square brackets. It may be possible to construct diabolical cases where there is some ambiguity (such as for unpaired square bracket characters), but no such cases have been identified in any existing dataset and it is expected to be vanishingly rare that these diabolical cases would correspond to real data files. The rationale is that we expect the user community to construct USIs themselves and it is unrealistic to expect them to insert escape characters.

A secondary use case of this element is for individual repositories to distinguish between different instances/processing of the original MS runs. This does detract from the universal aspect of USI, but is potentially quite useful for interactions with a single repository. For example, for PXD015028 (ftp://massive.ucsd.edu/MSV000084207/), there are two different versions of HELA-DDA-1.mzML (one with MS1 scans in centroid-mode and another with MS1 scans in profile-mode), and the [subFolder] notation MAY be used to disambiguate between them. If a USI with a subFolder prefix is sent to a repository where there are no subfolders available for a particular dataset, it may choose to ignore the subFolder, ideally with a warning to the user.

3.6.2 Loss of original scan number in plain text formats

As of this writing, the use of MGF and other plain text formats remains common practice, despite their substantial loss of information. A common loss of information in MGF files is the loss of the original scan number from the raw file. This loss is not inherent to MGF but is common with many MGF converters. The official MGF specification (http://www.matrixscience.com/help/data_file_help.html) does provide the RAWSCANS and SCANS keywords to preserve this information. Use of these keywords would lead to greater preservation of provenance information

. In some software implementations the scan number is encoded in the TITLE.

In order to compensate for this problem (while encouraging the community promote better practices), USI provides for an alternate indexing scheme. Instead of scan:17555, the same spectrum when sourced from an MGF file might be index:8104, where index:8104 represents the 8104th spectrum in the MGF beginning with 0. MGF files typically do not have an explicit unique index number but rather just have an implicit Nth spectrum. This is obviously not desirable, but perhaps a necessary patch for the problem while better solutions become widely used.

MassIVE MGF USI example using index:
mzspec:PXD007592:good_responder_1_2.mgf:index:22627:TLM+15.994915TQIDGVNLAANSLVESGHPR/3

MassIVE MGF USI example using scan number:
mzspec:MSV000081142:a5fe923d686b45928cad9a1f24567c05.mgf:scan:8694:MIAETSSGGVAAN+0.984016DVIVHITLHSLPFGGVGNSGMGSYHGK/3

Additionally, a common occurrence with MGF files is that original scans are combined into a single spectrum. A record of exactly which scans were combined may or may not be available. If multiple scans are combined into a new spectrum, then that spectrum does not exist as a single original scan and cannot be referenced by scan numbers - the only way to access it is using its index.

3.6.3 Which instance of the scan/spectrum?

The primary idea for the USI is to try to refer (for the most part) to a specific data generation event from an instrument, somewhat independent of what happens to it downstream. So, ideally a USI will refer to a specific scan event in a raw file/folder, although different repositories may have different views or instances of that scan. Imagine scan 1000 in a Thermo RAW file that is submitted to PRIDE. That scan 1000 represents an acquisition that is the evidence for some peptide ion. The original authors may have run Distiller on it that did some deisotoping and charge reduction (hopefully preserving the original scan number) and analyzed that with Mascot. PeptideAtlas may have downloaded the RAW file and created an mzML file with centroiding using msConvert (doi.org/10.1038/nbt.2377). MassIVE may have downloaded that RAW file and implemented software that can directly read the scan in profile mode from the RAW file using Thermo’s MSFileReader/ RawFileReader code. Each repository will have a somewhat different instance of that instrument scan event. The primary intent of the USI is to refer to that original scan event and whatever comes downstream. Imagining a request for that USI to the three different repositories, in each case, the exact form of the spectrum returned may be a little different (different post processing), but fundamentally, if it is a high quality spectrum, then the ultimate identification ought not depend on what processing it has undergone (although differences in the computed precursor *m*/*z* and reference database used can, in practice, lead to incomplete/incorrect/different PSM assignments). Ultimately, the concept is that an instrument somewhere digitized a signal from a biological entity and it is that digitized event that we want to refer to. If different analysis paths of that original event lead to different interpretations, that is something we can argue about, but the most fundamental concept is what that instrument generated.

However, there is interest in also using the USI to refer to specific downstream instances of that scan event. This probably comes at a loss of universality, but with the possible benefit in interacting with individual repositories. For example, MassIVE may have several instances of a particular MS run (e.g. a .wiff file, an mzXML file produced with msConvert, an mzML file produced with SCIEX's converter, another mzML file that has undergone mass recalibration and denoising). These different encodings may be specified by placing a filename suffix at the end of the MS Run. While references to a .RAW file are universal, it is understood that a .mzML suffix at one repository may yield a different conversion than at another repository.

3.6.4 Use of nativeId instead of scan as an indexType

As mentioned above, the goal of the USI is to refer to an original scan event that generated a spectrum, and using the indexType "scan" is preferred. However, for some instrument types (most vendors other than Thermo Scientific), a single scan number cannot uniquely identify a spectrum, and instead a set of integers is required to identify a scan. This issue was solved in mzML (5) via the use of the nativeId mechanism. As an example, one scan event is identified in an mzML file converted from a SCIEX WIFF file with:

sample=1 period=1 cycle=2740 experiment=10

In this scenario, where reference to the original scan event is desired but a single scan number is not sufficient, the USI must be formed with a **compact form** of the nativeId mechanism: the tag "nativeId" MUST be placed in the indexType field, followed by a comma-separated set of integers that correspond to the full-length nativeId as indexNumber. Therefore, a USI employing this mechanism might look like:

mzspec:PXD001464:CL_1hRP_rep3:nativeId:1,1,2740,10

The number and order of the values is vendor specific and is defined by the nativeId controlled vocabulary terms in the PSI-MS controlled vocabulary as children of term MS:1000767 (http://purl.obolibrary.org/obo/MS_1000767). A few examples are provided below. See the CV for the full set:

SCIEX WIFF format (MS:1000767):
sample=1 period=1 cycle=2740 experiment=10 →nativeId:1,1,2740,10

Waters nativeId format (MS:1000769):
function=10 process=1 scan=345 →nativeId:10,1,345

Bruker TDF format (MS:1002818):
frame=120 scan=475 →nativeId:120,475

Thermo nativeId format (MS:1000768) SHOULD NOT be expressed as a nativeId, but rather as a scan:
controllerType=0 controllerNumber=1 scan=43920 → scan:43920
since the controllerType and controllerNumber are always 0 and 1 for mass spectra. In rare cases, if either controllerType is not 0 or controllerNumber is not 1 (e.g., a PDA spectrum is being referenced), then the nativeId form MUST be used:
controllerType=5 controllerNumber=1 scan=7 → nativeId:5,1,7
The use of the scan:43920 form means that controllerType=0 controllerNumber=1.

The order of the keys is crucial and must be ordered as defined in the PSI-MS CV nativeId format.

For example, the following USIs can be resolved by nativeId:

mzspec:PXD001587:18302_REP2_500ng_HumanLysate_SWATH_2.mzML:nativeId:1,1,2,2:HAVSEGTK

3.6.5 Data Independent Acquisition (DIA) Data

DIA data (such as SWATH-MS or similar on other platforms) present some slight additional challenges regarding USIs. Primary spectra can be referenced as already described based on the previously described scheme. They are, however, heavily multiplexed and therefore any single interpretation will only be one of many peptides represented in the spectrum. Multiple simultaneous interpretations MAY be signified by stringing together several interpretation strings with a plus (+) symbol, such as EMEVEESPEK/2+ELVISLIVER/3.

The more complex case is when DIA data are analyzed with a demultiplexing algorithm such as DIA-Umpire (7), wherein new pseudo-spectra are generated based on demultiplexing co-occurring peaks, ideally corresponding to a single precursor ion. The demultiplexed spectra are then typically analyzed with a conventional search engine such as Comet. If these demultiplexed spectra are submitted along with the original dataset (with modified MS run names to indicate non-original data), then whichever indexing scheme is supported by the format those files are in can be used for later USI referencing. However, if the demultiplexed spectra are written in a format that resists indexing, this may complicate the ability to reference them with USIs. Furthermore, if the original dataset is reprocessed and demultiplexed with a different tool, the filenames and index orders will be different, eliminating the ability to refer to the same peak group in a universal manner.

For example, in the original DIA-Umpire work, USIs referencing the original multiplexed scans look like this:

- mzspec:PXD001587:18300_REP2_500ng_HumanLysate_SWATH_1:scan:4974:EEAAEYAK
- mzspec:PXD001587:18300_REP2_500ng_HumanLysate_SWATH_1:scan:4974:GDSSAEELK
- mzspec:PXD001587:18300_REP2_500ng_HumanLysate_SWATH_1:scan:4974:M[+15.994915]SAEDIEK

whereas USIs to demultiplexed spectra would refer to differently named pseudo-runs or perhaps the same run names with a different subfolder, depending on how the submitter had organized the data. Demultiplexed spectra are available under different MS run names along with the original instrument files. For example, for peptide TLTPIIQEYFEHGDTNEVAEMLR

- USI for Pseudo-MS2 from 27854_LongSwath_EIF4aJune7-Biorep2_Q3.mzXML: mzspec:PXD001587:27854_LongSwath_EIF4aJune7-Biorep2_Q3.mzXML:scan:2072:TLTPIIQEYFEHGDTNEVAEMLR
    - Scan numbers are just indices of extracted pseudo-MS2 spectra
    - There is no 1-to-1 correspondence with original .wiff/.scan files
    - Link to original PSM search context
    - Link to PSM in dataset update
    - Originally in LongSwath_EIF4aJune7-Biorep2.wiff / .scan
- USI for mzML matching nativeId: mzspec:PXD001587:LongSwath_EIF4aJune7-Biorep2.mzML:nativeId:1,1,1431,22:TLTPIIQEYFEHGDTNEVAEMLR
- USI for mzXML matching scan number: mzspec:PXD001587:LongSwath_EIF4aJune7_Biorep2.mzXML:scan:50072:TLTPIIQEYFEHGDTNEVAEMLR

3.6.6 The PSM Provenance Identifier

In order for repositories to encode a globally unique identifier for the provenance of a particular instance of a PSM, one additional field is envisioned of the form:

mzspec:<collection>:<msRun>:<indexType>:<indexNumber>:<interpretation>:<provenanceID>

such as (not a working example):

mzspec:PXD000561:Adult_Frontalcortex_bRP_Elite_85_f09:scan:17555:VLHPLEGAVVIIFK/2:PR-G47

This final suffix begins with a two-letter repository code, a minus character, followed by an arbitrary string (ideally short) unique to the repository that provides a unique identifier for a particular PSM. There are several reasons why a given repository may have several instances of a PSM for a single spectrum. For example, multiple searches of the same spectrum could yield the same interpretation but at different false discovery rates because of search context or search parameters. This suffix thus enables repositories to disambiguate between multiple searches yielding the same PSM. It can also be used to determine at which repository one could look to examine full provenance information for a PSM. The range of allowed repository codes is defined at https://github.com/HUPO-PSI/usi/blob/master/RepositoryCodes.md.

Repositories other than the one that is responsible for creating a particular PSM Provenance Identifier (determinable via the two-letter prefix) must be prepared to handle such a USI, but it is not stipulated here how they are handled. The receiving resource might reject the request as being related to a foreign identifier; or it might display to the user what it can provide about the spectrum itself, ignoring (ideally with a warning to the user) the fact that it is an identifier from a foreign repository; or it might open a connection to the originating repository and fetch information about that PSM Provenance Identifier to provide to the user.

For example, if a user were to copy-paste the above PSM provenance identifier into the PeptideAtlas web interface, the interface may locate the referenced spectrum mzspec:PXD000561:Adult_Frontalcortex_bRP_Elite_85_f09:scan:17555, annotate it with the suggested interpretation VLHPLEGAVVIIFK/2, and provide a prominent notice that states A) whether or not the suggested interpretation matches PeptideAtlas reprocessing results, and B) that the original provenance information behind the supplied PSM provenance identifier is stored at PRIDE, displaying a suitable hyperlink to take the user to the PRIDE web interface with the PSM provenance identifier to read more information about the original context of the PSM.

This feature has been designed, but not implemented yet anywhere as of this time.

3.6.7 Authentication

Ideally the use of USIs would be supported during the manuscript review process so that reviewers could click or copy paste USIs for important claims. This would require some sort of authentication mechanism (or security through obscurity) for viewing spectra in datasets that are not yet public. The consensus is that credentials MUST NOT be embedded in the USI, but rather, a separate https-based authentication scheme be advocated.

3.7 Implementations Notes

There is an important distinction between the USI, which is just an identifier, and how it will be used in PROXI and other web services. Implementation details are not included here in this document. Up-to-date information will be maintained at http://psidev.info/USI and https://github.com/HUPO-PSI/usi/blob/master/Implementations.md.

4 Authors Information

Eric W. Deutsch
Institute for Systems Biology, Seattle WA, USA
edeutsch@systemsbiology.org

Juan Antonio Vizcaíno
European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, United Kingdom
juan@ebi.ac.uk

Yasset Perez-Riverol
European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, United Kingdom
yperez@ebi.ac.uk

Jeremy Carver
University of California San Diego, San Diego CA, USA
jjcarver@ucsd.edu

Shin Kawano
Database Center for Life Science, Kashiwa, Japan
kawano@dbcls.rois.ac.jp

Pierre-Alain Binz
CHUV Centre Universitaire Hospitalier Vaudois, CH-1011 Lausanne 14, Switzerland
pierre-alain.binz@chuv.ch

Benjamin Pullman
University of California San Diego, San Diego CA, USA
bpullman@eng.ucsd.edu

Jim Shofstahl
Thermo Fisher Scientific | 355 River Oaks Parkway | San Jose | CA 95134 | USA
jim.shofstahl@thermofisher.com

Ralf Gabriels
Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium
VIB-UGent Center for Medical Biotechnology, Ghent, Belgium
Ralf.Gabriels@UGent.be

Tim Van Den Bossche
Department of Biomolecular Medicine, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium
VIB-UGent Center for Medical Biotechnology, VIB, Ghent, Belgium
Tim.VanDenBossche@UGent.be

Nuno Bandeira
University of California San Diego, San Diego CA, USA
bandeira@ucsd.edu

5 Contributors

In addition to the authors, a number of additional contributions have been made during the preparation process. The contributors who actively participated to the recommendation documentation are:

- Matt Chambers, Stamford, CT, USA
- Henry Lam, Hong Kong University of Science and Technology, Hong Kong SAR China
- Zhi Sun, Institute for Systems Biology, Seattle, WA, USA
- Luis Mendoza, Institute for Systems Biology, Seattle, WA, USA
- Gerben Menschaert, Ghent University, Ghent, Belgium
- Yunping Zhu, Beijing Proteome Research Center, Beijing, China
- Wout Bittremieux, University of California, San Diego, CA, USA
- Tytus Mak, National Institute of Standards and Technology, Gaithersburg, MD, USA
- Joshua Klein, Chestnut Hill, MA, USA
- Mingxun Wang, University of California, San Diego, CA, USA

6 Intellectual Property Statement

The PSI takes no position regarding the validity or scope of any intellectual property or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; neither does it represent that it has made any effort to identify any such rights. Copies of claims of rights made available for publication and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the PSI Chair.

The PSI invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights which may cover technology that may be required to practice this recommendation. Please address the information to the PSI Chair (see contacts information at PSI website).

7 Copyright Notice

Copyright (C) Proteomics Standards Initiative (2021). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction of any kind, provided that the above copyright notice and this paragraph are included on all such copies and derivative works. However, this document itself may not be modified in any way, such as by removing the copyright notice or references to the PSI or other organizations, except as needed for the purpose of developing Proteomics Recommendations in which case the procedures for copyrights defined in the PSI Document process must be followed, or as required to translate it into languages other than English.

The limited permissions granted above are perpetual and will not be revoked by the PSI or its successors or assigns.

This document and the information contained herein is provided on an "AS IS" basis and THE PROTEOMICS STANDARDS INITIATIVE DISCLAIMS ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE

INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE."

8 Glossary

There does not appear to be a need to have a glossary since all non standard terms are already defined in detail in section 3.

9 References

1. Bradshaw, R. A., Burlingame, A. L., Carr, S., and Aebersold, R. (2006) Reporting protein identification data: the next generation of guidelines. *Mol. Cell Proteomics* 5, 787–788

2. Chalkley, R. J., MacCoss, M. J., Jaffe, J. D., and Röst, H. L. (2019) Initial Guidelines for Manuscripts Employing Data-independent Acquisition Mass Spectrometry for Proteomic Analysis. *Mol. Cell Proteomics* 18, 1–2

3. Deutsch, E. W., Overall, C. M., Van Eyk, J. E., Baker, M. S., Paik, Y.-K., Weintraub, S. T., Lane, L., Martens, L., Vandenbrouck, Y., Kusebauch, U., Hancock, W. S., Hermjakob, H., Aebersold, R., Moritz, R. L., and Omenn, G. S. (2016) Human Proteome Project Mass Spectrometry Data Interpretation Guidelines 2.1. *J. Proteome Res.* 15, 3961–3970

4. Bradner, S. (1997) RFC2119: Key words for use in RFCs to Indicate Requirement Levels (https://tools.ietf.org/html/rfc2119).

5. Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W. H., Römpp, A., Neumann, S., Pizarro, A. D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., and Deutsch, E. W. (2011) mzML--a community standard for mass spectrometry data. *Mol. Cell Proteomics* 10, R110.000133

6. Wimalaratne, S. M., Juty, N., Kunze, J., Janée, G., McMurry, J. A., Beard, N., Jimenez, R., Grethe, J. S., Hermjakob, H., Martone, M. E., and Clark, T. (2018) Uniform resolution of compact identifiers for biomedical data. *Sci Data* 5, 180029

7. Tsou, C.-C.; Avtonomov, D.; Larsen, B.; Tucholska, M.; Choi, H.; Gingras, A.-C.; Nesvizhskii, A. I. DIA-Umpire: Comprehensive Computational Framework for Data-Independent Acquisition Proteomics. Nat. Methods 2015, 12 (3), 258–264, 7 p following 264. https://doi.org/10.1038/nmeth.3255.