PubChem Substance Deposition System allows the use of SD formatted data files. The SD file format is described elsewhere and will not be described here. To learn more about the SD file format, go to: http://www.mdli.com/downloads/public/ctfile/ctfile.jsp

Provided below is documentation on the required and allowed SD fields to be used during deposition of PubChem Substance data. Those SD fields not matching those in this documentation will be ignored.

Please note that the SD fields may evolve over time as we modify, change, or revise the PubChem Substance Deposition System.

PubChem Deposition Required SD Field

PUBCHEM_EXT_DATASOURCE_REGID

Depositor's Unique External Registry ID. This external registry ID or name is provided by you, the depositor. It is your primary key to a substance and ***must be unique*** across all data ***deposited by you***, the data source. If you provide an External Registry ID that is not unique to your depositions, it will be treated as an update request that will replace the existing substance record in PubChem with the data you provide in the SD record.

Only a single registry ID is allowed in a single line of data for this SD field. The expected format must be visible ASCII text without control characters. Rich Text Format (RTF), HTML codes, e.g., "<I>", and HTML4 special characters, e.g., "ο", are strongly discouraged. External Registry IDs with disallowed characters may prevent proper deposition of the Substance.

PubChem Deposition Allowed SD Fields

PUBCHEM_EXT_DATASOURCE_SMILES

SMILES string to be used to represent the chemical structure for the Substance being deposited. This is an ***alternate mechanism*** for providing a structural description of a Substance and will be ignored if a chemical structure with atoms is also provided in the SD file format CTAB section in the same SDF record deposited.

Only a single SMILES string is allowed for a given Substance. The expected format is a single line of text containing a valid SMILES string. Please ***Kekulizé your SMILES*** (i.e., no aromatic atoms in the SMILES), otherwise, the chemical structure could be ambiguously interpreted.

PubChem Deposition Allowed SD Fields (cont.)

PUBCHEM_PUBLICATION_DATE

PubChem Publication Date. Date deposited data is to be made available on the public PubChem data system. While we will attempt to honor this release date to the best of our abilities, we may disallow deposition of substances with release dates greater than two months.

Only a single date is to be provided on a single line of text. The expected format is modeled after the international standard date notation ISO 8601 and must be one of the following:

Complete date:
YYYY-MM-DD (e.g., 1997-07-16)
Complete date plus hours and minutes:
YYYY-MM-DDThh:mmTZD (e.g., 1997-07-16T19:20+01:00)

where:

YYYY = four-digit year
MM = two-digit month (01=January, etc.)
DD = two-digit day of month (01 through 31)
hh = two digits of hour (00 through 23) (am/pm **NOT** allowed)
mm = two digits of minute (00 through 59)
TZD = time zone designator (Z or +hh:mm or -hh:mm)

There are two independent ways of handling time zone offsets:

[1] Times are expressed in UTC (Coordinated Universal Time), with a special UTC designator ("Z").

[2] Times are expressed in local time, together with a time zone offset in hours and minutes. A time zone offset of "+hh:mm" indicates that the date/time uses a local time zone which is "hh" hours and "mm" minutes ahead of UTC. A time zone offset of "-hh:mm" indicates that the date/time uses a local time zone which is "hh" hours and "mm" minutes behind UTC.

Examples:

1994-11-05T08:15-05:00
corresponds to November 5, 1994, 8:15 am, US Eastern Standard Time

1994-11-05T13:15:30Z
corresponds to the same instant in time as the prior example

PubChem Deposition Allowed SD Fields (cont.)

PUBCHEM_EXT_DATASOURCE_INCHI

InChI string to be used to represent the chemical structure for the Substance being deposited. This is an ***alternate mechanism*** for providing a structural description of a Substance and will be ignored if a chemical structure with atoms is also provided in the SD file format CTAB section in the same SDF record deposited.

Only a single InChI string is allowed for a given Substance. The expected format is a single line of text containing a valid InChI string.

PUBCHEM_DEPOSITOR_RECORD_DATE

Depositor's Record Date. This optional field allows you, the depositor, to specify the internal creation or modification date of your substance record. This is not intended to be used as a deposition or export date; rather it is intended to be the date the substance was last changed in your internal database (maps to Entrez index field "SourceReleaseDate"). Please note that PubChem automatically provides (within Entrez) a date when the record is added or updated (Entrez index field "DepositDate").

Only a single date is to be provided on a single line of text. The expected format is identical to the PUBCHEM_PUBLICATION_DATE field.

PUBCHEM_EXT_DATASOURCE_URL

Depositor's External Source/Database Universal Resource Locator (URL). An URL may be provided to link back to your, the depositor's, web site from within PubChem. This URL will be associated with your external source/database name for the Substance deposited.

Only a single external source/database URL is allowed for a given Substance. The expected format is a single line of text containing the URL exactly as it is required for use in the PubChem web pages.

PUBCHEM_EXT_SUBSTANCE_URL

Depositor's External Substance Universal Resource Locator (URL). An URL may be provided to link back to your, the depositor's, web site in PubChem. This URL will be associated with your external registry ID/name.

Only a single external registry ID/name URL is allowed for a given Substance. The expected format is a single line of text containing the URL exactly as it is required for use in the PubChem web pages.

PubChem Deposition Allowed SD Fields (cont.)

PUBCHEM_REVOKE_SUBSTANCE

Revoke Substance Record. Use of this field will notify PubChem to suppress this Substance from the Entrez search system. The Substance will remain in the PubChem archive; however, there will be no direct links to this substance within the PubChem system.

Only a single line of text is allowed with a comment about why this Substance is to be revoked.

PUBCHEM_SUBSTANCE_COMMENT

Substance Comments. Comments are your, the depositor's, textual annotations about the Substance. Comments may be used to provide beneficial information about a Substance record that can be found through exact or keyword text searches. We reserve the right to suppress or eliminate unsuitable or excessive comments.

Multiple Comments are allowed for a Substance. The expected format must be printable ASCII characters. HTML4 special characters will be automatically removed. URL's via HTML "<A>...</A>" may be used.

PUBCHEM_SUBSTANCE_SYNONYM

Substance Synonyms. Synonyms allow alternate ways to identify, link, and locate the provided Substance. Synonyms may include common names, systematic names, and trade names. Synonyms help provide a way for users to find your Substance record through exact or keyword text searches.

Providing useful and multiple synonyms for a substance will increase the chance that your substance will be located by users of PubChem. By you providing synonyms and a valid chemical structure, you will additionally help integrate your substance with other Entrez databases.

Multiple Synonyms are allowed for a Substance. The expected format must be ASCII text without control characters. HTML codes and HTML4 special characters will be removed. Only a single Synonym is expected per line.

PUBCHEM_PUBMED_ID

NLM/NIH PubMed ID for an article or abstract. PubMed ID's enables users to realize an association between the Substance record and the article associated with the PubMed ID and vice-versa. This must be a valid PubMed ID.

Multiple PubMed ID's are allowed for a Substance. The expected format is an unsigned number. Only a single PubMed ID is allowed per line.

PubChem Deposition Allowed SD Fields (cont.)

PUBCHEM_GENBANK_GENERIC_ID

NCBI/NLM/NIH GenBank General ID. GenBank ID's enable users to realize an association between the Substance record and a protein or nucleotide sequence via the GenBank ID and vice-versa. This must be a valid GenBank General ID (GI) and ***not a GenBank Accession ID***.

Multiple GenBank ID's are allowed for a Substance. The expected format is an unsigned number. Only a single GenBank ID is allowed per line.

PUBCHEM_BONDANNOTATIONS

Substance Bond Annotations. Bond Annotations allow you, the depositor, to provide additional structural information that may not be readily encoded in the SD file format. Bond Annotations will affect how the Substance is interpreted and validated within PubChem.

Multiple Bond Annotations may be provided for a Substance. The allowed format for a Bond Annotation is three unsigned numbers, separated by white-space, per line, representing the AtomID's of the two atoms, followed by the annotation ID, respectively. Only a single Bond Annotation may be provided per line. The atoms do not have to be explicitly bonded in the SD file format to have a bond annotation. Nonsensical annotations will be suppressed.

Atom-Atom Annotation list is in the format:
AtomID AtomID AnnotationID

where AtomID and AnnotationID are unsigned integer numbers.

AnnotationID	Meaning
1	Crossed Bond, a non-specific stereo double bond
2	Dashed Bond, a 3-D hydrogen bond
3	Wavy Bond, a non-specific stereo single bond
4	Dotted Bond, a complex or fractional bond
5	Wedge-up Bond, a solid wedge stereo bond
6	Wedge-down Bond, a dashed wedge stereo bond
7	Arrow Bond, a dative bond
8	Aromatic Bond, an aromatic bond
9	Resonance Bond, a resonating bond
10	Bold Bond, a thick bond
11	Fischer Bond, use Fischer stereo conventions
12	Close Contact, a 3-D atom-atom close contact

PubChem Deposition Allowed SD Fields (cont.)

PUBCHEM_NCBI_TAXONOMY_ID

NCBI/NLM/NIH Taxonomy ID. Taxonomy ID's enable users to realize an association between the Substance record and an NCBI Taxonomy record and vice-versa. This must be a valid Taxonomy ID.

Multiple Taxonomy ID's are allowed for a Substance. The expected format is an unsigned number. Only a single Taxonomy ID is allowed per line.

PUBCHEM_NONSTANDARDBOND

Substance Non-Standard Bonds. Non-Standard Bonds allow you, the depositor, to provide additional information that may not be readily encoded in the SD file format. Non-Standard Bonds will affect how the Substance is interpreted and standardized within PubChem.

Multiple Non-Standard Bonds may be provided for a Substance. The allowed format for a Non-Standard Bond is three unsigned numbers, separated by white-space, per line, representing the AtomID's of the two atoms, followed by the bond type ID, respectively. Only a single Non-Standard Bond may be provided per line. The atoms do not have to be actually bonded in the SD file format to have a non-standard bond. If the atoms are already bonded in the SD file format, the non-standard bonds provided using this SD tag will supersede that interpreted from the SD file format.

Atom-Atom Non-Standard Bond list in the format:
AtomID AtomID BondTypeID

where AtomID and BondTypeID are unsigned integer numbers.

BondTypeID	Meaning
1	Single Bond
2	Double Bond
3	Triple Bond
4	Quadruple Bond
5	Dative Bond
6	Complex Bond
7	Ionic Bond

PubChem Deposition Allowed SD Fields (cont.)

PUBCHEM_NCBI_OMIM_ID
NCBI/NLM/NIH Online Mendelian Inheritance in Man (OMIM) ID. OMIM ID's enable users to realize an association between the Substance record and an NCBI OMIM record and vice-versa. This must be a valid OMIM ID.

Multiple OMIM ID's are allowed for a Substance. The expected format is an unsigned number. Only a single OMIM ID is allowed per line.

PUBCHEM_NCBI_MMDB_ID
NCBI/NLM/NIH MMDB ID. MMDB ID's enable users to realize an association between the Substance record and an NCBI MMDB record and vice-versa. This must be a valid MMDB ID.

Multiple MMDB ID's are allowed for a Substance. The expected format is an unsigned number. Only a single MMDB ID is allowed per line.

PUBCHEM_NCBI_GENE_ID
NCBI/NLM/NIH Entrez Gene ID. Gene ID's enable users to realize an association between the Substance record and an NCBI Entrez Gene record and vice-versa. This must be a valid Gene ID.

Multiple Gene ID's are allowed for a Substance. The expected format is an unsigned number. Only a single Gene ID is allowed per line.

PUBCHEM_NCBI_PROBE_ID
NCBI/NLM/NIH Entrez Probe ID. Probe ID's enable users to realize an association between the Substance record and an NCBI Entrez Probe record and vice-versa. This must be a valid Probe ID.

Multiple Probe ID's are allowed for a Substance. The expected format is an unsigned number. Only a single Probe ID is allowed per line.

PUBCHEM_GENERIC_REGISTRY_NAME
Substance Generic Registry Name/ID. Generic registry names are typically assigned by an outside organization. This must be a valid Registry Name or ID.

Multiple Registry Names or IDs are allowed for a Substance. The expected format is either an unsigned number or a series of three unsigned numbers delimited by a "-" character. Only a single Registry Name or ID is allowed per line.

How to Export an SD File from ISIS/Base for PubChem Deposition

1. Have your ISIS db file ready.
2. Load your ".db" file into ISIS/Base.
3. In 'Query' mode, click 'Search->Retrieve All', or import a list.
4. Click 'Database->View Definition…'.
5. In the View Database Definition window, choose a field, such as 'compound_id', then click 'Modify External Name…' button.
6. In the new window displayed with title 'Change external field name', enter the corresponding PubChem tag, such as 'PUBCHEM_EXT_DATASOURCE_REGID', and click 'OK'.
7. Repeat steps 5 and 6 until you finish setting the PubChem SD tags for all fields you desire to deposit into PubChem.
8. Click 'File->Export->SDFile…'.
9. Select the fields you intend to export into the SD file.
10. Choose a file name to save.

Please Note: The 'PUBCHEM_EXT_DATASOURCE_REGID' SD field, which reflects your unique external registry ID, **is required** for every Substance to enable PubChem deposition.

Document Version History

V1.4.1 – 2007Jul09 – Added new SD field PUBCHEM_NCBI_PROBE_ID.
V1.4.0 – 2007Mar06 – Added new SD field PUBCHEM_EXT_DATASOURCE_INCHI. Modified wording of PUBCHEM_EXT_DATASOURCE_URL and PUBCHEM_EXT_SUBSTANCE_URL.
V1.3.0 – 2006Mar23 – Added new SD field PUBCHEM_NCBI_GENE_ID. Numerous minor edits to the SD field descriptions.
V1.2.2 – 2005Dec08 – Added new SD field PUBCHEM_NCBI_OMIM_ID and modified wording of PUBCHEM_DEPOSITOR_RECORD_DATE.
V1.2.1 – 2005Jun10 – Modified wording of PUBCHEM_GENBANK_GENERIC_ID.
V1.2.0 – 2005May05 - Added two new SD fields PUBCHEM_DEPOSITOR_RECORD_DATE and PUBCHEM_REVOKE_SUBSTANCE. Added "Document Version History" section.
V1.1.0 – 2005Apr26 - Altered the PUBCHEM_PUBLICATION_DATE to allow for a specific time.
V1.0.0 – 2005Apr20 - Initial release.