Data quality and consistency

Control of errors, data origin and quality

An experimental measurement can be marked as an “error”. Such records are highlighted with a red background and indicate a possible problem. The system allows users to manually mark a record as an error if they believe there is a mistake. In this case, the user should provide an explanation of the problem in the comment or discussion field related to this record. The OCHEM system can also automatically mark records as erroneous if they do not comply with the system rules. Namely, a record is automatically marked as an error if:

- an obligatory condition of the experiment has not been specified (for example, a boiling point measurement without specifying the pressure is ambiguous and would be marked as an error automatically)
- a duplicate of the record exists in the database (see the next section for the definition of “duplicate”)

Another quality indicator is the “to be verified” flag. This flag signals that the record has been introduced from a referencing article, e.g., benchmarking/methodological article and should be verified against the original publication. This flag can be set either manually or automatically by the system (e.g., in case of batch data upload, see the “Batch upload” section for details).

Duplicates management

To ensure data consistency, it is essential to avoid redundancy in the database. Thus, there is a need for strict rules for the definition of duplicates. In OCHEM two experimental records of a physicochemical or biological property are considered to be duplicates if they are obtained for the same compound under the same conditions, had the same measured value (with a precision up to 3 significant digits) and are published in the same article. We refer to these records as *strong duplicates*, as opposed to *weak duplicates*, for which only part of the information is the same. The OCHEM database does not forbid strong duplicates completely, but forces all the duplicates (except for the record introduced first) to be explicitly marked as errors. This ensures that there are no strong duplicates among the valid (i.e., non-error) records.

The uniqueness of chemical compounds is controlled by special molecular hashes, referred to as InChI-Keys [19]. Namely, for the determination of duplicated experimental measurements, two chemical structures are considered the same if they have identical Inchi-keys.

OCHEM allows weak duplicates (for example, completely identical experimental values, published in different articles) and provides facilities to find them. Moreover, in the modeling process, it is always automatically ensured that the same compounds in the training set appear only in onefold of the N-fold cross-validation process.

Experimental data origin

Each record has a colored dot indicating the origin of the data. Green dots indicate “original records” from publications with a description of experimental protocols; these are usually the publications where the property was originally measured (original data). The users can verify experimental conditions and experiments by reading these articles. These are the most reliable records in the database. The weak duplicates of *original records* have magenta dots. The other records have red dots and originate from articles that re-use the original data but for which the original records are not stored. These are frequently methodological QSAR/QSPR studies. The original records can be easily filtered out by checking a corresponding box in the *compound property browser*. Another filter, “primary records”, eliminates all weak duplicates except the record with the most early publication date.