1. A section that describes the party owning and/or using the chemical substance. This section includes administrative information about the company including company contacts, company site information and geographic locations. This section also includes information on the volumes of the substance in question, by tonnage. Some of these attributes are free text data types, but the information is structured.

2. A section describing the substance itself. This section describes the substance/substance mixtures/compositions, by means of identifiers that correspond to international de facto standards, e.g. CAS Number, EC Number, IUPAC name, and molecular and structural formulae. Note that there could be multiple substances described e.g. in the case of a mixture or composition, in nested sections under section 2. Such subsections will have the same structure and format as the top-level section 2.

3. A section providing information about the substance, that we describe as endpoints, meaning reports of physiochemical properties, examinations, or blood tests, research reports or similar, which are unstructured.

The new data source is volatile in the sense that it is subject to changes in data structure, content and meaning.

The current volume of the new data source is ~ 500GB, and is expected to grow moderately, by ~15% year on year over the next 10 years. There is an average record size of ~150KB, but this can vary significantly from a few tens of KB to several tens of MB depending on the number and type of endpoint data included.

The data source provides as a service, events that signal when the data has been updated, once every 1000 records.

Requirements

- The new data source is to be made available through a common data store, which will make the new data source available to the data processing systems, considering the specific, differing needs of each component listed in section Background above.

- Considering the volatile nature of the data source, the solution should be robust, configurable and extensible over time

- The common data store should employ mechanisms to minimise the impact of changes in the data source in terms of the need to refactor the consuming systems of the data platform

- The common data store should be able to accommodate additional data sources in the future, which may feature NoSQL, relational, unstructured data stores

- The common data store should collect appropriate metadata on the data source and include mechanisms to make this available to the consuming systems of the data platform

- The solution proposed should employ appropriate security mechanisms to ensure access to the source data is restricted to the consuming systems of the ECHA data