11 November 2022

Pre-release

EU-SRS

Proteins EU-SRS User Guide

Guidance on naming and building protein records in EU-SRS

Disclaimer

This document is created as part of deliverable D2.8 EU-SRS Data Management Plan of Unicom Work Package 2: Implement IDMP – Substance Management in Europe.

This guide will be a living document, used by the Substances Validation Group (SVG) for creation and maintenance of substances in EU-SRS.

The current version is a pre-release. Your feedback, if any, is welcomed by **8 December 2022.** Comments can be sent to Steven de Wit (e-mail: s.d.wit@cbg-meb.nl).

Your feedback will be considered when preparing the official release which will be submitted as Unicom deliverable to the European Commission in January 2023.

Document control

This document is subject to a regular review by the Substance Validation Group (SVG). It is a living document, and changes will be captured in the version history section.

Document ownership

This document is owned by the SVG.

Revision history

Version	Date	Changes made	Author(s)
N/A	11.11.2022	Pre-release	Mirjam Keulen, Semra Olsthoorn

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of abbreviations

Abbreviation	Complete Form
BAN	British Approved Name
BP	British Pharmacopoeia
CAS	Chemical Abstracts Service
CV	Controlled Vocabulary
Da	Dalton
EDQM	European Directorate for the Quality of Medicines and HealthCare
EMA	European Medicines Agency
EP	European Pharmacopoeia
EPAR	European public assessment report
EUTCT	European Union Telematics Controlled Terms
EU-SRS	European Substance Registration System
EVMPD	EudraVigilance Medicinal Product Dictionary
FDA	Food and Drug Administration
G-SRS	Global Substance Registration System
IGG	immunoglobulin G, an antibody type
INCI	International Nomenclature Cosmetic Ingredient
INN	International Nonproprietary Name
INN PL	INN proposed list
INN RL	INN recommended list
ISO	International Organization for Standardization
ISO IDMP	ISO
ISO/TS	ISO Technical Specifications
JAN	Japanese Accepted Name
KEGG	Kyoto Encyclopedia of Genes and Genomes
MAB	monoclonal antibody
NCATS	National Center for Advancing Translational Sciences
NCI	National Cancer Institute
Ph. Eur.	European Pharmacopoeia
RMS	Referentials Management Service
PT	Preferred Term
rINN	recommended International Nonproprietary Name
SMS	Substance Management Service
SMSID	Substance Identifier assigned by EMA SMS

Abbreviation	Complete Form
SPOR	Substances, Products, Organisations, Referentials
SSG	Specified Substance Group
STN	Scientific and Technical information Network
SVG	Substance Validation Group
SVGID	leading identifier in EU-SRS
TNF	tumor necrosis factor
UNII	Unique Ingredient Identifier
USAN	United States Adopted Name
USP	United States Pharmacopeia
WHO	World Health Organization
xEVMPD	Extended EudraVigilance Medicinal Product Dictionary

1 Introduction

The EU Network is currently implementing the ISO IDMP standards in a phased programme based on the four domains of master data in pharmaceutical regulatory processes: substance, product, organisation and referential (collectively referred to as “SPOR”) master data. ISO IDMP compliant business services for the central management and supervision of data in each of the four SPOR areas will be established through an iterative and incremental delivery approach. Through the Substance Management Services (SMS) of the SPOR programme EMA will provide the EU network centralised substance data management services.

EU-SRS will become the scientifically rigorous back-end for the Substance Management Services of SPOR. EU-SRS will be accessible to the EU regulatory network, enabling the unambiguous identification of substances used in medicinal products based on their scientific properties in accordance with ISO IDMP standard 11238 and ISO IDMP technical specification standard 19844. EU-SRS allows the unique identification of substances which will support various purposes including the enhancement of traceability of pharmacovigilance, non-clinical, clinical and quality findings with a high degree of precision to substances by their scientific identity.

The Substance Validation Group is responsible for building substance records in EU-SRS. In addition, the SVG defines guidance and best practices for substances management in EU-SRS (per substance type).

1.1 Purpose

The purpose of this document is to provide practical guidance for the registration of proteins in EU-SRS.

1.2 Scope

Registration of human and veterinary proteins are in scope of this document.

This document is intended to be used together with the EU-SRS Substance Maintenance Process which describes the workflow between EMA and SVG (under development).

The document is currently written mainly based on monoclonal antibodies (MAB’s). In the future the document will be complemented with information of other protein types.

2 Defining a protein

In order to ensure good quality protein substance records built in a well-organized, and harmonised way, rules have been established and agreed upon within the SVG together with relevant partners. References to external documentation are made where necessary. This chapter provides general guidance on how to define proteins in EU-SRS.

The concepts required for the unique identification and description of substances are described in the ISO 11238 IDMP standard on substances. Guidelines for implementing ISO 11238 are provided in the technical specification ISO/TS 19844. Although ISO 11238 does not provide any guidance on substance nomenclature, it does provide a structure for the capture of names and codes that are used to refer to a substance. This section aims to provide supplementary guidance and should be read in conjunction with the standard and technical specification.

2.1 Protein

Peptides of 3 or more amino acids and proteins are described as protein in EU-SRS. A peptide consists of one subunit with a defined linear sequence of 2-40 alpha-amino acids connected through peptide bonds. A protein consists of one subunit with a defined linear sequence of more than 40 alpha-amino acids connected through peptide bonds or multiple subunits with a defined linear sequence of alpha-amino acids connected through peptide bonds. See Figure 1 for an overview of all proteinogenic amino acids, amino acids that are incorporated biosynthetically into proteins during translation. Figure 2 shows the formation of a peptide bond.

Defining elements of peptides and proteins are the amino acid sequence(s) of their subunit(s), disulfide bond(s), site(s) and type of glycosylation and other essential modification(s). The production method (i.e. synthetic, recombinant, purified from biological matrix) is not a defining element. So proteins that differ in amino acid sequence(s), disulfide bond(s), site(s) or type of glycosylation or other essential modifications get a separate protein record in EU-SRS.

Sequence. Monopeptides, dipeptides, cyclic peptides, non-ribosomal peptides (e.g. cyclosporines, bleomycins and vancomycin) and peptides that consist mostly of non-proteogenic amino acids are described as chemical in EU-SRS. Mixtures of proteins with different sequences are described as mixture or structurally diverse substance in EU-SRS (e.g. polyclonal immunoglobulins). Proteins are defined by the final expressed sequence and not as pro-protein. The letter X is used to represent amino-acids in synthetic peptides that are not part of the 22 proteinogenic amino acids (Figure 1). Proteins are described as different substances when their sequences differ. Interferon alfa-2a and interferon alfa-2b differ at a single residue and are therefore described as 2 different substances.

Non-glycosylated peptides and small proteins get a chemical substance record as alternative definition to capture the molecular structure of the protein.

Type of glycosylation. The site and type of glycosylation is captured at substance level. The type of glycosylation describes the cell or organism where the protein was isolated from or produced (e.g. fungal, plant, anthropoid, avian, mammalian and human). The type of glycosylation reflects significant differences in overall glycosylation. Glycosylation in human and old-world monkeys differs for example from other mammals in lacking the alpha, 1,3-galactose epitopes. In humans and primates terminal sialic acid residues are only acetylated and not glycolated. Coagulation factor VII isolated from human blood and coagulation factor VII produced by recombinant technology in Chinese hamster ovary (CHO) cells get separate records due to differences in glycosylation type. Trastuzumab and trastuzumab beta are produced in different cells and get separate records due to differences in glycosylation type.

Microheterogeneity, differences in the extent of occupancy and type of glycan present on a given site, is not captured at substance level. A glycoprotein produced in different mammalian cell lines is described with the same substance record in EU-SRS. Detailed information about glycosylation (occupancy, type of glycans) is captured at Specified Substance Group 1 level.

Modifications. Proteins often undergo extensive but variable post-translational modifications either within the cell or bioreactor or during processing, purification or storage of the protein. Extensively modified peptides are described as chemical in EU-SRS. Post-translational modifications are described at substance level when they are both complete and non-variable or when they are essential for bioactivity. For example the conversion of glutamate to gamma-carboxyglutamate in many blood clotting factors is essential for activity and therefore captured as structural modification in EU-SRS. Proteins with essential differences in post-translational modifications are described as different substances in EU-SRS. For example human serum albumin and aggregated human serum albumin, formed by irreversible partial physical denaturation, get separate protein records.

Microheterogeneity, slight differences in the structure of proteins, is not captured at substance level. Post-translational modifications are only captured in detail when the protein is intentionally modified or when essential for the functionality of the protein.

Production method. Recombinant and chemically synthesised salmon calcitonin are chemically identical and therefore considered to be the same substance. Somatropin, a non-glycosylated protein, is defined as the same substance regardless of production in E. coli, yeast or mammalian cells. Vaccines and toxins (e.g. botulinum toxin) consisting of protein subunits are described as structurally diverse substances. A protein record is part of the hierarchy.

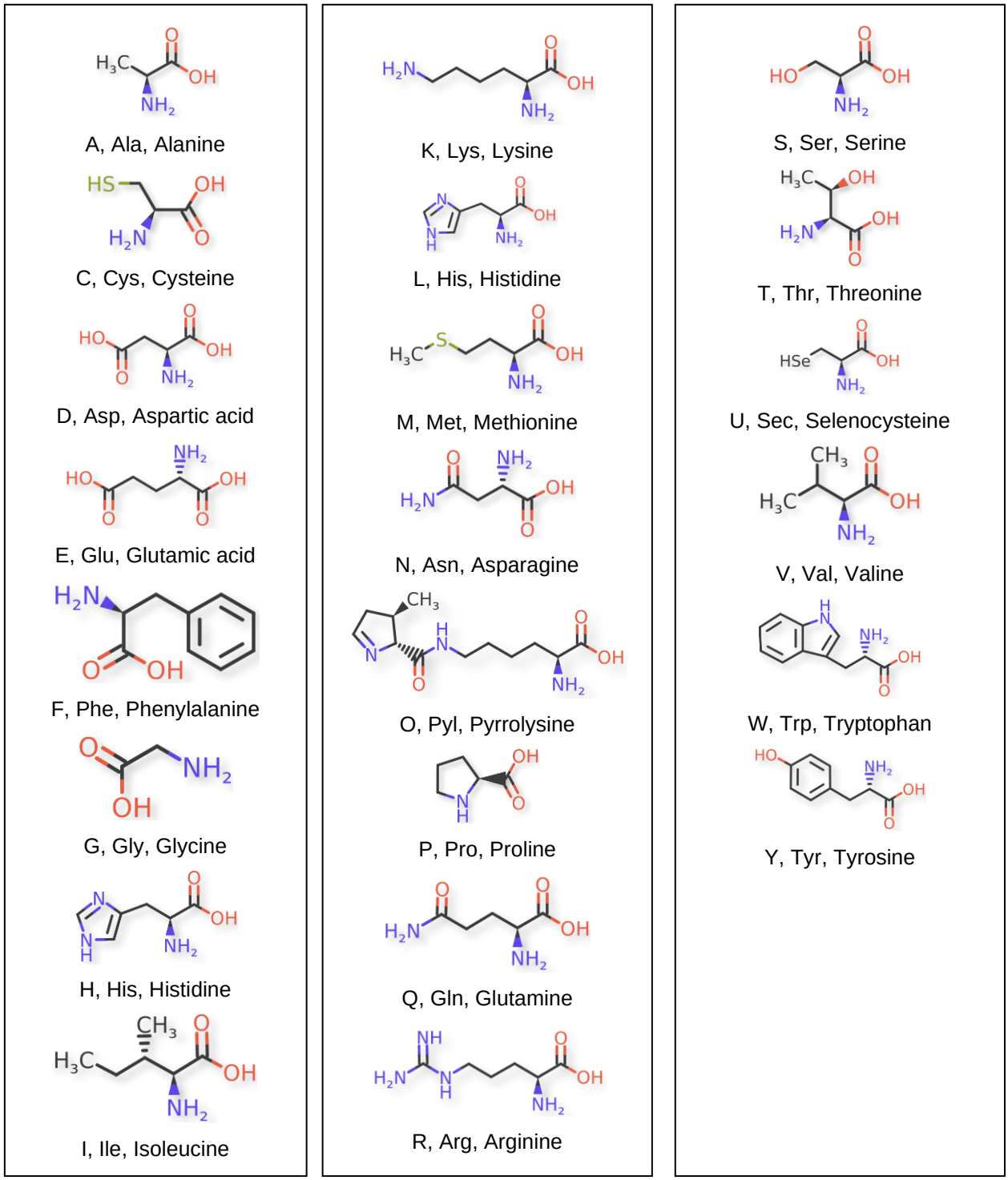


Figure 1. Proteinogenic amino acids, amino acids that are incorporated biosynthetically into proteins during translation

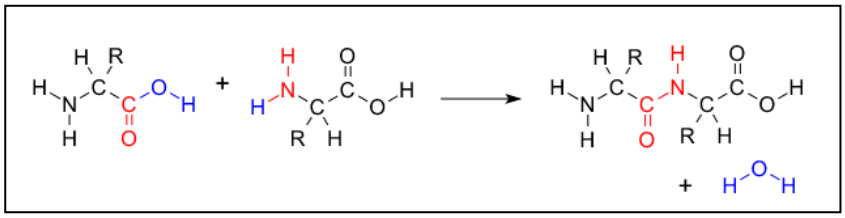


Figure 2. Formation of a peptide bond

2.2 Protein type

Protein types currently recognized are:

(NB: this list might change in the future)

- Enzyme
- Fusion protein
- Monoclonal antibody
- Monoclonal antibody conjugate
- Monoclonal antibody fusion protein
- Peptibody
- Peptide
- Protein
- Protein, Radiolabelled
- Vaccine antigen
- FAB
- FAB2

2.3 Protein subtype

Protein subtypes currently recognized are:

(NB: this list might change in the future)

- Allergen
- Anticoagulant
- Bone Morphogenetic Protein
- Calcium Channel Blocker
- Carrier Protein
- Cell Adhesion Protein
- Chemokine
- Coagulation Factor
- Collagen
- Colony Stimulating Factor
- Cytokine
- Cytochrome P450
- Cytokine Receptor
- Enzyme Inhibitor
- Esterase
- Glucosidase
- Gonadotropin
- Growth Factor
- Hemagglutinin
- Hormone
- Hydrolase
- Hyaluronidase

- IGG1
- IGG2A
- IGG2B
- IGG3
- IGG4
- Immunomodulator
- Insulins
- Interferons
- Interleukins
- Lipase
- Neuramidase
- Neurotrophin
- Neuropeptide
- Neurotoxin
- Nuclease
- Parathyroid
- Phosphatase
- Protease
- Receptor
- Stem Cell Factor
- Sulfatase
- Sulfohydrolase
- TNF Binding Protein
- Toxin
- Toxin Conjugate
- Tumor Necrosis Factor
- Viral Capsid

3 Naming of Proteins

This chapter provides details around naming of proteins in EU-SRS and SMS.

3.1 Name types

EU-SRS contains name type information. ISO 11238 lists several name types. Applicable name types in EU-SRS for proteins are:

- **Brand names:** Names set by the marketing company.
- **Codes:** Synonyms that represent company codes.
- **Common names:** Usage for first designation of recognised trivial names that have an additional systematic designation. An alternative for "official name".
- **Official names**: This name type contains comprehensive information about the name. It is chosen for names that are assigned by an official organisation. According to the EU Directive, the International Nonproprietary Name (INN) and the European Pharmacopoeia must be followed in the EU.
- **Systematic names:** Names that describe the structure of a protein.

3.2 Naming convention

Each unique substance receives an SMSID from SMS and each SMSID has a Preferred Term, which is characterised as 'Display Name' in the system. The Preferred Term is the name most accurately describing the substance at a given time and could change during the lifetime of a substance.

The Preferred Term is used in several forms visible to industry. Going forward, it is planned that these forms (such as eAF) will also display the aliases.

3.2.1 Preferred term

The Preferred Term of a protein should be selected according to the priority ranking of the following reference sources and name types:

1. European Pharmacopoeia (Ph. Eur.) (Official Name Type)

   Exception: Ph. Eur. name with pharmaceutical product characteristics as part of the substance name is not valid as preferred name or alias. Example: Alteplase for injection.

   Exception: Species specific proteins (see 3.2.2)
2. Recommended International Nonproprietary Name (rINN) (Official Name Type)

   Exception: Species specific proteins (see 3.2.2)

   The INN experts have adopted a general scheme for the naming of peptides and proteins, available at https://www.who.int/teams/health-product-and-policy-standards/inn/stembook
3. Other official name type with EU jurisdiction (e.g. INCI) (Official Name Type)

   Exception: Species specific proteins (see 3.2.2)
4. Common name (Common Name Type)

Recommendations:

- The amount of salts and hydrates should not be included in the name unless necessary to differentiate between different salts/hydrates of the same protein.

  Example: 'Terlipressin acetate' is preferred name of terlipressin diacetate.
- The production method should not be included in the name when it is not defining for the protein (sequence, disulfide bonds, type and sites of glycosylation, essential modifications).
- Choose for proteins without an official European name that belong to a group of proteins that have an official European name a common name that is in line with the official names.

Example: Interferon alfa-2b (common name) is preferred name. This name is in line with the INN names of other interferons e.g. Peginterferon alfa-2a and peginterferon alfa-2b.

- When naming proteins which can be grouped into a family based on homology or according to a notion of shared function (like the interleukins), the different members should be enumerated with a dash '-' followed by an Arabic number. Example: desmoglein-1, desmoglein-2, etc.
- A recommended name should not contain a Roman numeral.

  Example: caveolin-2 is preferred to caveolin-II

  Exception: historical cases. e.g. coagulation factor IX, casein kinase II, HLA class I, etc.
- Abbreviations should not be built using the molecular weight.

  Example: Abbreviations such as p123, Gp62, p34 are not suitable.

  Exception: cases where historically the molecular weight has been consistently and generally applied as part of the accepted name, e.g. p53.
- A name that ends with '-in' is preferred over a name that ends with '-ine'.

  Example: zyxin, insulin, hemoglobin, caveolin, desmoglein, secretin, etc.
- See 3.2.2 for the naming rules of species specific proteins.

5. Company code (Code Name Type)

3.2.2 Naming rules for species specific proteins

For species specific proteins the official Ph. Eur. or INN name should not be used, because they are not consistent. Instead, the following rules should be used:

1. Plasma-derived proteins (according to Ph. Eur.): species – protein (name of the species at front)

   Examples:
   - Human C1-esterase inhibitor (Ph. Eur.)
   - Human coagulation factor VII (Ph. Eur.)
   - Human fibrinogen (Ph. Eur.)
2. Other proteins: protein – comma – species

   Examples:
   - Thrombin, bovine (Ph. Eur.)
   - Calcitonin, salmon (Ph. Eur. name is Calcitonin (salmon))
   - Secretin, human (INN is Secretin human)
3. When a protein can be produced by multiple species, specify the name of the species

   Examples:
   - Corticorelin, human (Corticorelin is INN)
   - Corticorelin, ovine
   - Calcitonin, human (Calcitonin is INN)
   - Calcitonin, salmon (Ph. Eur. name is Calcitonin (salmon))
4. When there is only a human form of the protein, the species doesn't have to be mentioned

   Example:
   - Fibronectin (no official name available)
5. Salts and hydrates: name of the salt/hydrate after the protein name

   Examples:
   - Calcitonin acetate, salmon
   - Secretin pentahydrate, human
6. When the official name is not used as the preferred name, the official name becomes an alias

7. When a protein is plasma-derived, add an alias with ‘plasma-derived’ in the name
   Example:
   - Human C1-esterase inhibitor, Plasma-derived (alias)
   - Human coagulation factor VII, Plasma-derived (alias)
   - Human fibrinogen, Plasma-derived (alias)

3.2.3 Aliases

Aliases are valid alternative names for a Preferred Term, according to valid reference sources. SMS provides aliases when available. The EU-SRS Preferred Term should be written in European English. Any US English term is however to be kept as an alias.

Valid aliases:

- Official name (Official Name Type)

Examples: Ph. Eur., INN, INCI

- Systematic name (Systematic Name Type)

Source can be for example CAS Scifinder and/or recommended INN list.

Example: Gamma1 heavy chain (Homo sapiens VH-IGHG1*03) (221-214’)-disulfide with kappa light chain (Homo sapiens V-KAPPA-IGKC*01); (227-227’’:230-230’’)-bisdisulfide dimer (description of canakinumab in recommended INN list).

- Product name (Brand Name Type)

Example: Ilaris (Brand name of canakinumab (INN))

- Latin Ph. Eur. or INN name (Official Name Type)
- Company code (Code Name Type)
- JAN name with ‘(genetical recombination)’ in the name (Official Name Type)

In the substance naming conventions of the Japanese Pharmacopeia the term 'genetical recombination' is the common part of the substance name for all recombinant substances.

Example: Palmiteplase (genetical recombination) (JAN) is also known as palmiteplase (INN). Palmiteplase is defined as a recombinant modified human tissue plasminogen activator. Therefore, the recombination is an integral part of the substance name. In this case Pamiteplase (genetical recombination) should be considered as a synonym of palmiteplase sharing the same SMSID.
Example: The INN nonacog alfa is defined as recombinant human coagulation factor IX therefore Nonacog alfa (genetical recombination) (JAN) is a synonym of nonacog alfa.

4 Building protein records

The protein record captures information about the protein (sub)type (e.g. monoclonal antibody of IgG1 subtype), target(s), names and identifiers, amino acid sequences of all subunits, disulfide bonds, glycosylation type and glycosylation sites, structural modifications, molecular formula and molecular weight and references.

Glycosylation details (occupancy, type of glycans) are captured at SSG1 level (See chapter 0).

Note that this chapter is currently written mainly based on Monoclonal Antibodies.

4.1 Duplicate check

The first step of any registration should be to ensure that the substance is not already in EU-SRS**,** to prevent from adding a duplicate. To ensure there are no duplicates, different elements can be used for verification, such as names, identifiers and amino acid sequences. The approach for duplicates check should be determined case-by-case. If duplicates are found, further investigation is required to determine if it's a true duplicate.
There are several ways to search for protein substances in EU-SRS: Global Search, Advanced Search and Sequence Search.

4.2 General information

Each substance is unique in EU-SRS. Select "Protein" from the "Register" menu. The protein registration form will be displayed (see Figure 3). Section cards are collapsible to ease navigation, however the elements in this form are:

- Overview
- Names
- Protein Details
- Subunits
- Other Links
- Disulfide Links
- Glycosylation
- Agent Modifications
- Structural Modifications
- Physical Modifications
- Codes
- Relationships
- Notes
- Properties
- References

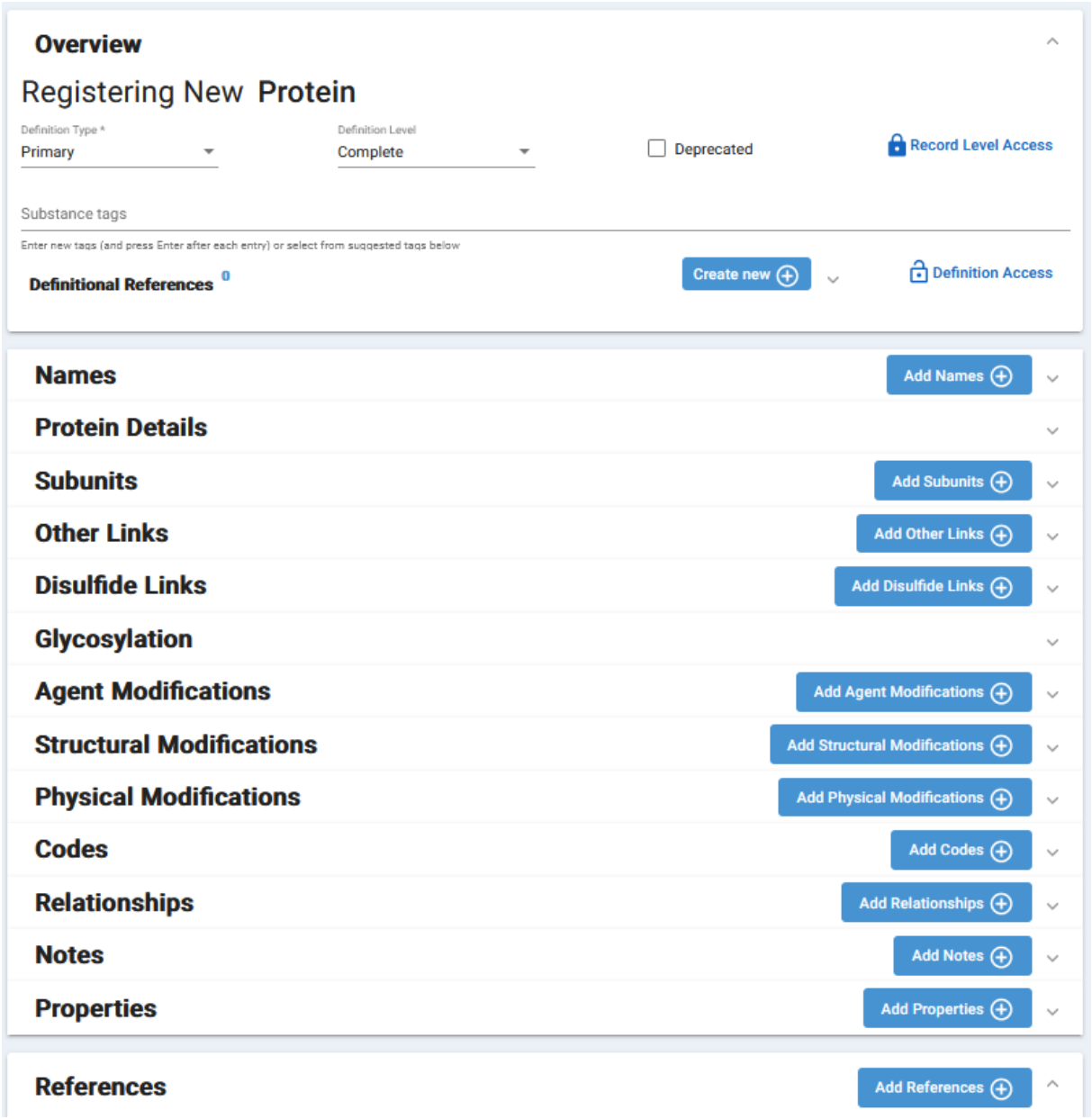


Figure 3. A form for registering a new Protein in EU-SRS

Mandatory fields for proteins are:

- at least one name with an unprotected reference marked “Public Domain” in order to be made public.
- at least one definitional reference
- at least one Subunit element (amino acid sequence) in order to be marked as complete.

After registration of the protein the following information is added automatically:

- the identifier “SVGID” with the reference “System Generated Code”
- the estimated molecular formula (based on the amino acid sequence)
- the estimated molecular weight (based on the amino acid sequence)
- validation messages in the section ‘Notes’
- the references “GSRS System-generated Validation messages’ and ‘System Generated Code”
- information about the date of registration and editing and the name of the person who registered and edited the record in the section ‘Audit Information’

4.3 Overview

4.3.1 Overview section

For the registration of a new protein in EU-SRS, some general information needs to be entered first (see Figure 4). Additional explanation of the EU-SRS fields is included in EU-SRS field guidance for registering and/or editing a protein.

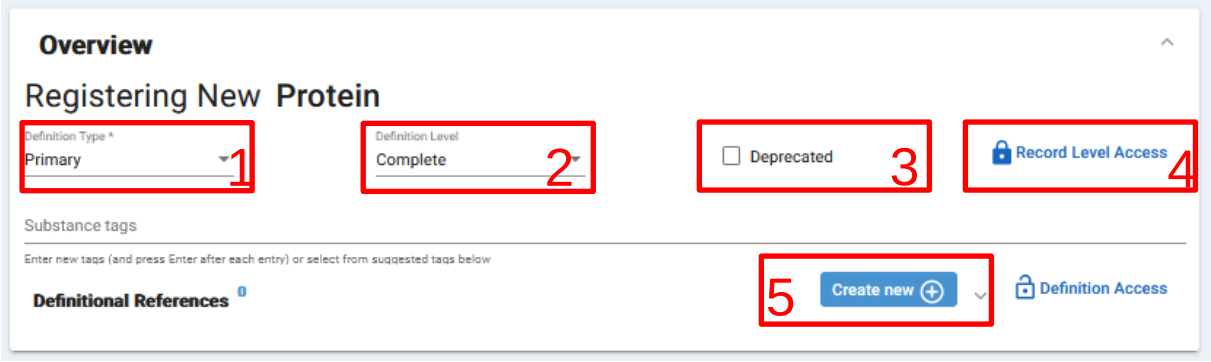


Figure 4. Overview of general information to be registered.

1. Definition Type - hover over and select the down arrow to activate the Edit drop-down. There are two options:
    a. Primary
    b. Alternative - Protein alternative definition (for example of structurally diverse substance). Once Alternative is selected a Primary Substance Search field appears

Note: Type the Primary Substance name in search box and select. As you begin, typing a list containing those letters will appear. Make a selection and then select search and the Primary Substance will populate. After submission, the system generates a relationship connecting both definitions. Alternative definition registration is similar to Primary definitions, but Names and Codes are not included.

2. Definition Level – select the down arrow to activate the Edit drop-down.
    a. Always try to enter a Complete definition (i.e. add an amino acid sequence of at least one subunit)
    b. Incomplete are allowed for incompletely defined substances.
    c. Representatives are too complex to define completely
3. Deprecated box: selecting this box means it is semi-deleted, or a candidate for deletion. There are other flags for deprecation elsewhere in EU-SRS, but that one also tells the browse/search functions not to show this record unless you specifically click the box allowing it to show.
4. Record Level Access: some substances are public in which case set the Record Level Access to Public by deselecting Protected under Record Level Access. Some substance are private in which case the option Protected should be selected. Private means not readily available in public sources, domains, website, etc. (e.g., SciFinder, PubChem).
5. Definitional Reference(s):
    a. Select the Create new + button
    b. The Add Reference screen appears and additional fields will be displayed

Note: Select the down arrow next to Source Type. Use the scroll on the right to navigate and select a Source Type. Source Text/Citation – identifies where the information comes from. Based on the release sensitivity: Check Public Domain if public (If public, add Tags for Public-Domain-Release). Update the Access by deselecting PROTECTED, if applicable. Select Save.

In order to make something Public it has to be made public three times:

- Undo the lock

- Select the Public Domain checkbox of at least one of the definitional references
- Enter the Tag “Public Domain Release” for at least one of the definitional references

Public Domain checkbox should be added for all public references. Every public name should have at least one reference with Public Domain checkbox. Public Domain Release Tag should be used for public records and added to the record when the lock is opened.

Click on Create new + to add more references or select a previously used reference by clicking on reuse to select a previous substance reference.

4.3.2 Overview of proteins

When registering a protein complete the Overview section:

- Definition type is ‘Primary’. Definition type is only ‘Alternative’ when there is also another type of record. For example ‘Clostridium botulinum, type A, strain Hall, neurotoxin complex’ has a structurally diverse record as primary record and a protein record as alternative record.
- Definition level is ‘Complete’ when the complete amino acid sequence is known.
- Definitional Reference. For technical reasons at least 1 reference with the tag ‘PUBLIC DOMAIN RELEASE’ must be added as definitional reference. This is the reference EMA List. Other definitional references for proteins are ‘European Procedure Number’ and ‘INN-list’ with citation ‘INN recommended List No. xx’ (in case of an INN).
- Undo the lock by deselecting the checkbox ‘PROTECTED’ of Record Level Access.

4.4 Names

4.4.1 Preferred names in EU-SRS and SMS

Names can be added with "Add Names +". Depending on the type, additional information is displayed. The language of the names is set to English by default. At the same time, the name is always present as "public".

For the correct mapping of the Preferred Term in SMS versus the Preferred Term in EU-SRS, the correct ticking of Display Name (DN) is crucial.

Option 1: SMS Preferred Term and EU-SRS Preferred Term (= Display Name) are the same

- In EU-SRS, the Preferred Term is selected as Display Name, and the Additional Listing Name is also ticked

Option 2: SMS Preferred Term and EU-SRS Preferred Term (= Display Name) are not the same

- In EU-SRS:
  - Preferred term is indicated as Display Name
  - SMS Preferred Term is indicated as Additional Listing Name
- In SMS:
  - The public term is indicated as Preferred Term
  - The EU-SRS Preferred Term is indicated as alias, with name source “Substance Validation Group”

Note that there are several reasons why in SMS the Preferred term is not the EU-SRS Preferred Term:

- The action to change the Preferred Term is postponed due to technical reasons
- The EU-SRS Preferred Term is a confidential term, and can therefore not be marked as Preferred Term; in SMS the PT is always a public term

Table 1 describes the different situations of correct mapping of the PT in SMS versus EU-SRS.

Table 1. Correct mapping of PT in SMS versus in EU-SRS.

Situation	SMS		EU-SRS	
	Name	Name Reference	Display Name (DN)	Additional Listing Name (AL)
SMS PT = EU-SRS PT	SMS PT = EU-SRS PT	-	Yes	Yes
SMS PT ≠ EU-SRS PT	Alias SMS = EU-SRS PT	Substance Validation Group	Yes	No
	SMS PT = EU-SRS Alias		No	Yes

4.4.2 Official name

The name of the substance is entered under Name, then the name type is selected under Type. Languages ("Languages": English) and Jurisdiction are selected from the drop-down lists.

Figure 5. Registering an Official Name.

Protein records can include more than one Official name, but just one name can be used as Display name (DN). The status of the official name is associated with the given jurisdiction. It is possible that an official name is a valid name in EU jurisdiction and other official name will be used in e.g. US jurisdiction.

For the official name, "Naming Organization" should be added (Figure 6). This is done with the "plus" sign next to "Naming Organization" when clicking on „Naming Organization“. If it is selected, the corresponding organization (BAN, EP, INCI, INN, JAN, USAN) can be selected. If more than one organization is to be registered, additional "Naming Organizations" must be added using the "plus" sign.

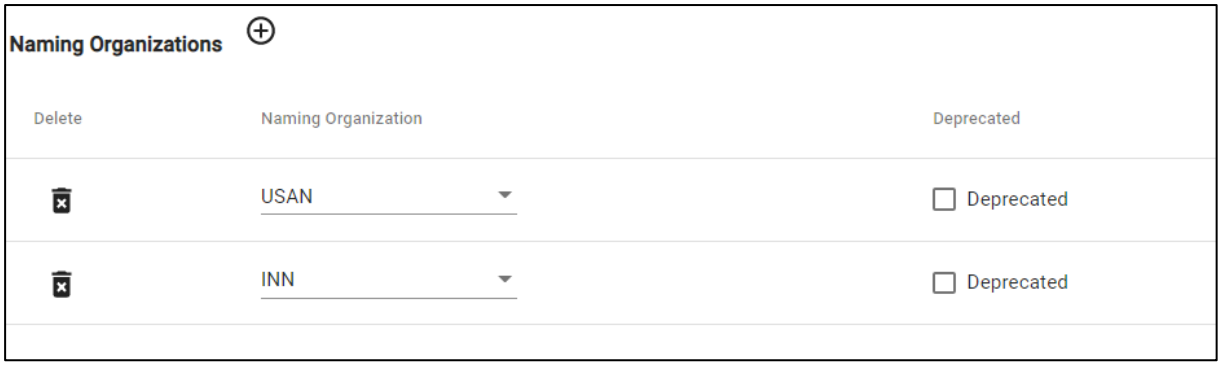


Figure 6. Registering Naming Organizations.

For Official Names it is mandatory to add the Jurisdiction to the relevant Naming Organization. This always corresponds to the countries for which this designation can be used. Table 2 shows the naming organizations that can be used for official names. It also shows how to correctly include the Jurisdiction when the Official Name is included as the Display Name or as an Alias.

Table 2. Listing of naming organizations and corresponding jurisdictions.

Naming organization	Description	Jurisdiction if Official Name is included as	
		Display Name	Alias
EP	European Pharmacopeia	EU	-
INN	International Nonpriority Name	EU	EU
USAN	United Sates Adopted Name	US*	US
USP	United States Pharmacopeia	US*	US
BAN	British Adopted Name	EU	GB
BP	British Pharmacopeia	EU	GB
JAN	Japanese Adopted Name	JP**	JP
INCI	International Nomenclature of Cosmetic Ingredients	EU	-

**If USAN/USP is the same as INN: then substance will get two Naming Organizations:*

1. INN (Jurisdiction EU)

2. USAN (Jurisdiction US)

If USAN/USP is not the same as INN, then INN will be included as DN (Jurisdiction EU), USAN/USP will be included as Alias (Jurisdiction US)

***If JAN is the same as INN: then substance will get 2 Naming Organizations:*

1. INN (Jurisdiction EU)

2. JAN (Jurisdiction JP)

If JAN is not the same as INN, then INN will be included as DN (Jurisdiction EU), JAP will be included as Alias (Jurisdiction JP)

Subsequently, the references for the corresponding name must be entered (Figure 7). References can be added with "Create new +". A dialogue window opens for this purpose. Using "add reference", the information on the source type (see the list of "Source types" used at the end of the document) and the "Citation" (free text for the actual citation) can be inserted. A checkmark should be placed next to "Public Domain" for the public names.

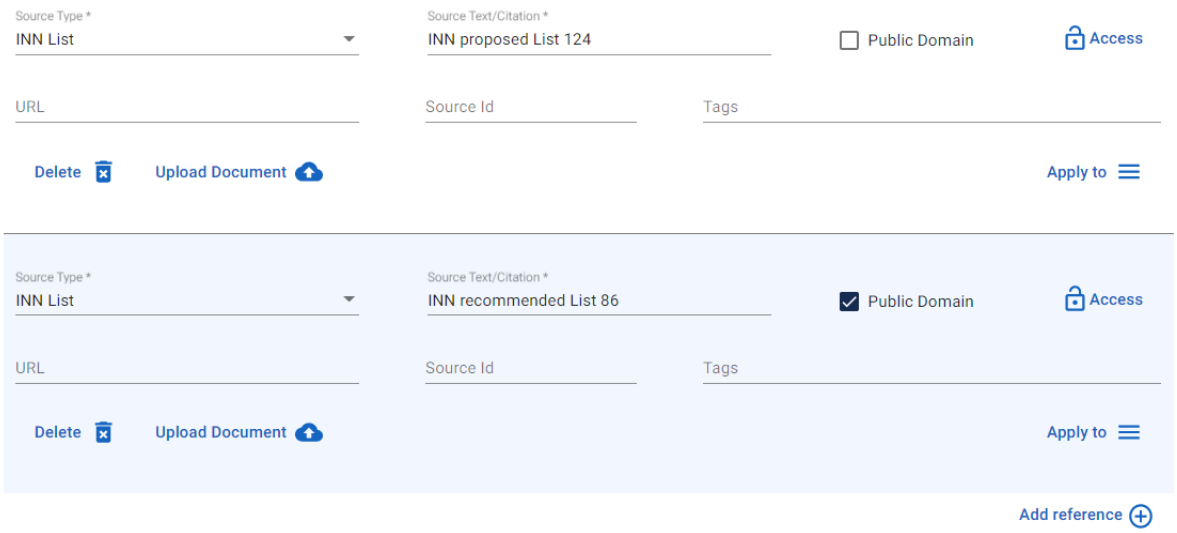


Figure 7. Registering references.

4.4.3 Systematic Name

The systematic name of the substance is entered under the tab "Name", then the name type is selected under "Type". As with the first name, the references are opened and saved in a new dialogue window ("reuse") (Figure 8).

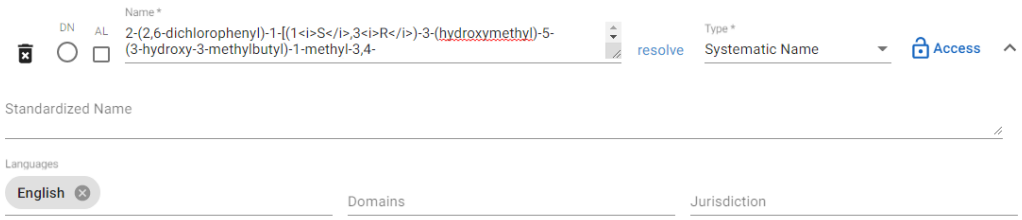


Figure 8. Registering of systematic names using HTML codes.

The systematic names from the recommended INN list and CAS Scifinder are used for proteins. The relevant references are attached.

The use of HTML meta-language ensures the correct use of Greek letters, Roman numerals, superscript and subscript, among others. The list of HTML codes to be used can be found in HTML code elements for displaying special characters.

4.4.4 Alias

Under the tab "Name" the name of the substance is added, then the name type is selected under "Type". English "Languages" is set by the program. No tick must be set for Display Name „DN“ or Additional Listing Name „AL“. The company codes can be entered with different spellings, depending on the reference database used (Figure 9).

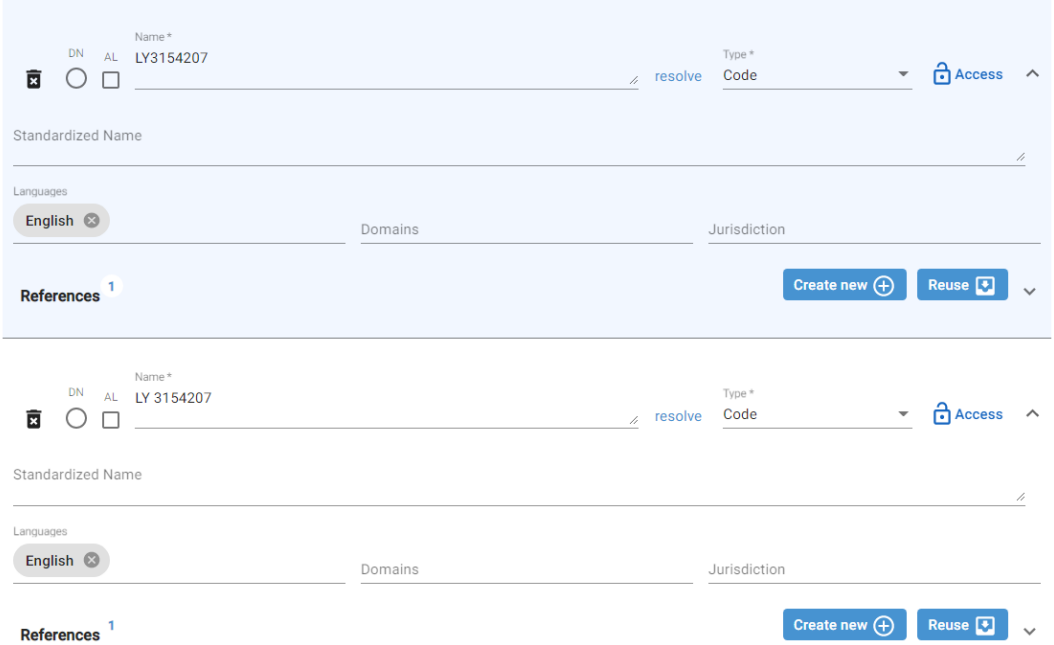


Figure 9. Registering company code with different spelling.

For the locked synonyms (protected, non-public), the access lock must be closed, and in the options under the lock symbol, a tick is set for "Protected".

4.4.5 Names of proteins

Names must be written in small letters with a capital at the beginning. Greek letters, Roman numerals, superscript and subscript must be written in HTML meta-language (see HTML code elements for displaying special characters). At least 1 public reference must be added to each name. The check box 'Display Name' is used for the preferred name in EU-SRS. The check box 'Additional Listing Name' is used for the preferred name in SMS.

In the names section fill out the following names:

- Preferred Name (see 3.2.1 Preferred term)
  Each record must have at least a preferred name. This is the display name. Check the box for 'Display Name'
  For example the INN with name type 'Official Name' and reference to the INN list with citation 'INN recommended List No. xx' (see 4.14 References).
  For an official name you have to complete the fields 'Naming Organizations' (e.g. 'INN', 'JAN', 'USAN') and 'Name Jurisdictions (e.g. 'European Union', 'United States'). See Table 2 for an overview of naming organizations and jurisdictions of proteins. The field 'Applicable Domains' must stay empty. The domain is covered with RMS codes in the code section.
  NB. When the Japanese Accepted Name is the same as INN, but with the addition of '(genetical recombination), JAN must not be added as naming organization.
- If INN is available: Systematic name(s) as published in the recommended INN-list.
  Type is 'Systematic Name'. Reference is INN list with citation 'INN description' (see 4.14 References).
  Proteins have often two descriptions in the INN list, separated by a semicolon. They must be added as two separate names to the record.
  Example: For canakinumab with this description in the INN list: "*immunoglobulin G1, anti-[Homo sapiens interleukin 1, beta (IL1B)] human monoclonal ACZ885; gamma1 heavy chain (Homo sapiens VH-IGHG1*03) (221-214')-disulfide with kappa light chain (Homo sapiens V-KAPPA-IGKC*01); (227-227'':230-230'')-bisdisulfide dimer*" the following systematic names are added to the record: '*Immunoglobulin G1, anti-[Homo sapiens interleukin 1, beta (IL1B)] human monoclonal ACZ885*' and '*Gamma1 heavy chain (Homo sapiens VH-IGHG1*03) (221-214')-*

*disulfide with kappa light chain (Homo sapiens V-KAPPA-IGKC*01); (227-227":230-230")-bisdisulfide dimer*'.

- Systematic name from CAS SciFinder
  Type is 'Systematic Name'. Reference is CAS with citation 'CAS SciFinder' (see 4.14 References).
- Product name (of originator)
  Type is 'Brand Name'. Reference is usually EPAR with citation 'EPAR' (see 4.14 References).
- Optional: other name(s) that give more understanding about the protein.

NB. In EU-SRS a name can only be used for 1 substance, so all names must be specific and unique.

4.5 Protein details

Complete the protein details:

- Protein Type. See chapter 2.2 for the protein types.
- Protein Subtype. See chapter 2.3 for the protein subtypes.
  Note: It is possible to select multiple protein subtypes
- Sequence Origin is for example 'human'.
- Sequence Type is 'complete' if the full sequence is known.

4.6 Subunits

Add the amino acid sequences of all subunits. The agreement is to start with the longest subunit and to end with the shortest subunit.

4.7 Disulfide links

All intra- and interchain disulfide bonds should be added.

The subunit and the amino acid with the smallest number is the first one of each pair. For example 1_22-1_96 describes a disulfide bond between amino acid 22 and 96 of subunit 1. For example 1_220-3_125 describes a disulfide bond between amino acid 220 of subunit 1 and amino acid 125 of subunit 3.

4.8 Glycosylation

The Glycosylation Type (e.g. mammalian) and the sites of N-, C- and O-glycosylation should be added.

4.9 Structural modifications

All posttranslational modifications (except glycosylation) with an amount > 1% should be added to the structural modifications section.

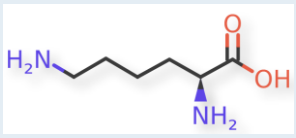
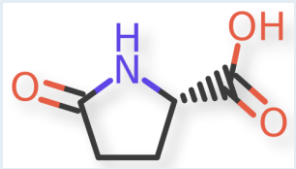
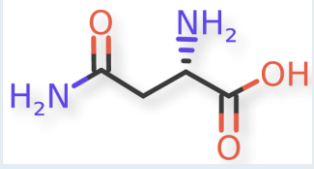
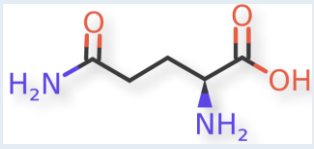
For each modification the following information should be added:

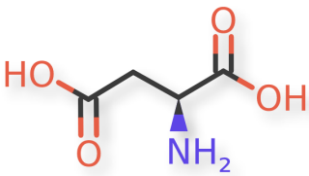
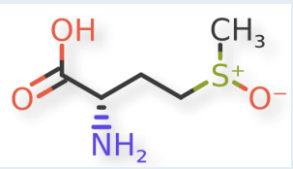
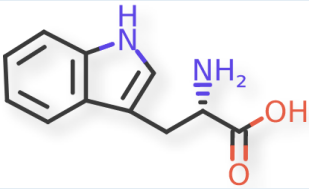
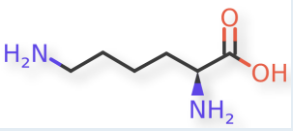
- Modification Type
  See Table 3 for an overview of common structural modifications.
- Fragment (after change)
  It is obligatory to fill in a fragment. Otherwise it is not possible to submit the record.
  If only one fragment is formed, add the formed fragment (e.g. pidolic acid is the fragment of N-terminal pyroglutamate formation).

If multiple fragments are formed or if the formed fragment is unknown, add the modified residue as fragment (e.g. asparagine as fragment of asparagine deamidation). See Table 3 for an overview of common structural modifications.

- Location Type
  The options are 'C-TERMINUS', 'N-TERMINUS', 'SITE SPECIFIC', 'RESIDUE SPECIFIC' AND 'UNKNOWN'. Select 'Site specific' when the exact location is known and select 'Residue specific' if only the residue is known and not the exact location.
- Sites of Residue Modified
  Choose the correct site(s) in the sequence(s) (when the location type is 'C-terminus', 'N-terminus' or 'Site specific') or add the modified residue (when the location type is 'Residue specific').
- Extent
  Choose between 'partial' and 'complete'.
- Extent Amount.
  Add the amount when the extent is 'partial'.
- Group
  Enter 1.

Table 3. Overview of common structural modifications

Modification	Modification Type	Fragment	Location Type
C-terminal lysine cleavage	C-terminal lysine cleavage	L-Lysine K3Z4F929H6 	C-terminus
N-terminal pyroglutamate formation	N-terminal pyroglutamate formation	Pidolic acid SZB83O1W42 	N-terminus
Asparagine deamidation	Asparagine deamidation	Asparagine 5Z33R5TKO7 	Site-specific
Glutamine deamidation	Glutamine deamidation	Glutamine 0RH81L854J 	Site-specific
Aspartate isomerization	Aspartate isomerization	Aspartic acid 30KYC7MIAI	Site-specific

Modification	Modification Type	Fragment	Location Type
			
Methionine oxidation	Methionine oxidation	Methionine sulfoxide XN1XVI4B2C 	Site-specific
Tryptophan oxidation	Tryptophan oxidation	Tryptophan 8DUH1N11BX 	Site specific
Glycation	Glycation	L-Lysine K3Z4F929H6 	Residue-specific

4.10 Identifiers

4.10.1 Identifiers and classifications

The code section in edit mode includes substance identifiers from other databases, RMS category and RMS domain. In browse mode all those codes will be displayed in two different sections. Substance identifiers from other databases will be displayed as identifier. RMS category and RMS domain will be shown as classifications.

4.10.2 Substance identifiers from other databases

The codes of the substance can be selected from the code system drop down list. Type of the code is also defined by the drop-down list. There are several types that can be selected from the list:

- Primary: to be used if the code uniquely describes the structure
- Generic (family): is selected if the code is not precisely specified and can be assigned to several substances, whole family
- Superseded: is set if the code numbers is already deprecated
- Alternative: is used when there are two codes that correctly describe the same structure. It is often found on SciFinder for recently recorded substances.

To enter the CAS No., CAS is selected from the “Code System” drop-down list (Figure 10). The type is set as described above. Cas No. is then inserted under Code.

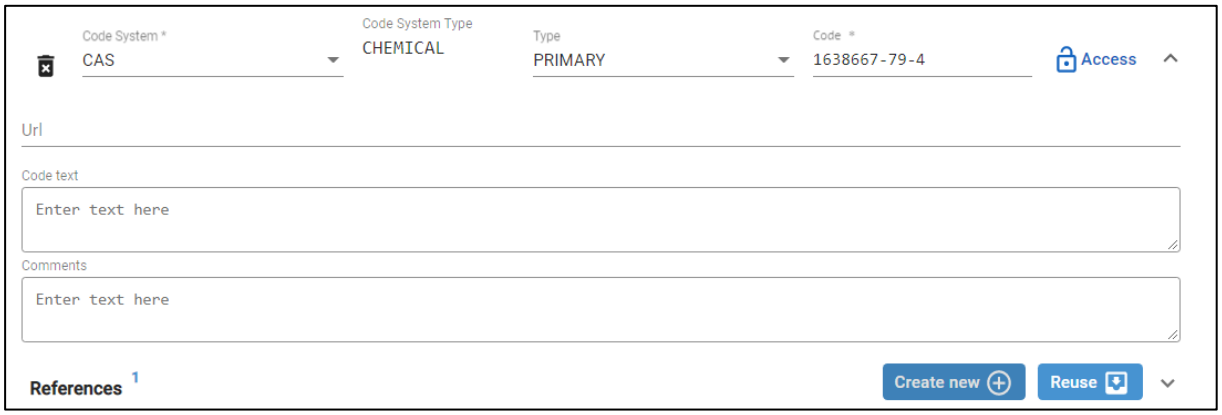


Figure 10. Registration of CAS ID.

CAS No. is a single identifier that requires a reference. Figure 11 shows how to enter a reference for CAS ID.

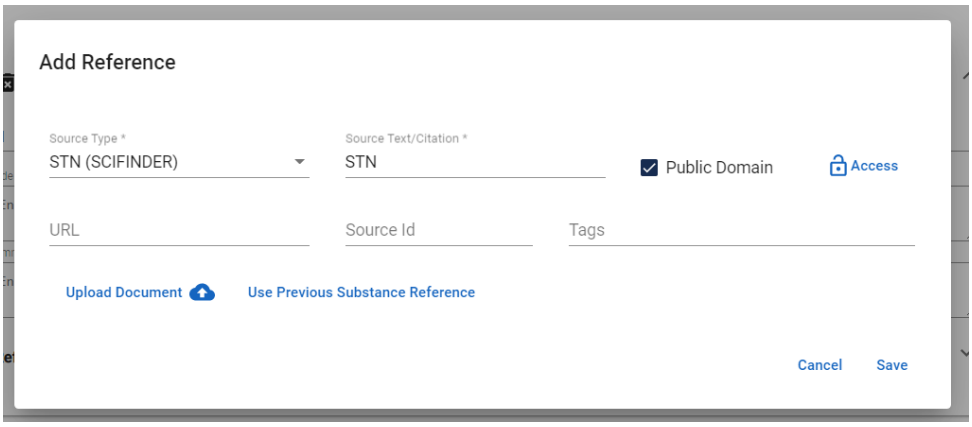


Figure 11. Reference for CAS ID.

Other codes (e.g. FDA-UNII, SMSID, xEVMPD, INN, EDQM) are included in the same manner, but they do not require a reference. EU-SRS will automatically construct URLs for the following Code Systems which can be accessed in the hyperlink on the view page. An overview of code systems and associated hyperlinks is included in Table 4.

Table 4. Code systems and associated hyperlinks.

Code System	Hyperlink
CAS	ChemIDplus (from 2023: CAS Common Chemistry)
FDA-UNII	G-SRS
PubChem	PubChem
INN	The School of INN
EDQM	Knowledge Database EDQM
NCI Thesaurus	NIH National cancer institute

4.10.3 RMS category and domain

The classification of the substance is entered under code section. To enter the “category” or ”domain”, RMS is selected from the “Code System” drop-down list. The type is set to “Generic (family)” and the number of the category is added under “Code”.

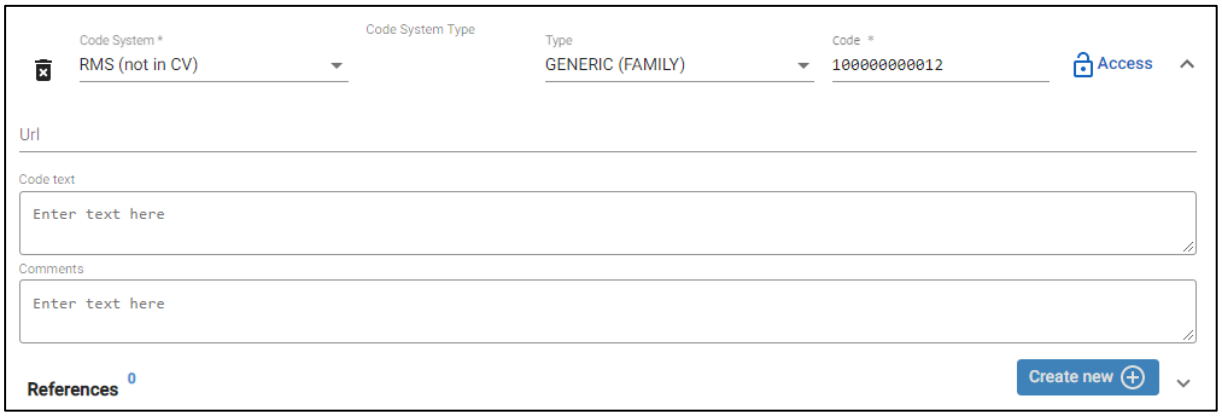


Figure 12. Registration of RMS domain.

Neither category nor domain need a reference. The use of the type generic is mandatory.

Some CVs have overlap with SMS and are managed in RMS. These CVs include:

- **Domain**: to indicate for which domain the substance is used for. This can be either human use or veterinary use (Table 5)
- **Category**: to indicate the substance type in EU-SRS (Table 5)

Table 5. Applicable RMS codes for Domain and Category for proteins.

RMS Code	Domain
100000000012	Human use
100000000013	Veterinary use
RMS Code	**Category**
200000005020	Protein

4.10.4 Identifiers of proteins

The identifiers section is used for codes.

This section can also be used for references, because it is possible to add hyperlinks that are automatically updated when the link changes. (NB. This is not functional in the training instance of EU-SRS.) The hyperlinks in the reference section can't be updated automatically.

It is also possible to add notes to the codes section. Notes in the identifiers section are more readable than notes in the notes section.

In the edit mode this section is called 'Codes'.

For each identifier the following fields should be added:

- Code System: See Table 6 for an overview of common identifiers used for proteins.
- Type: usually 'primary'.
- Code: See Table 6 for the codes that belong to each code system. The code is shown in the record and also used to automatically create hyperlinks. (NB. This is not functional in the training instance of EU-SRS.)
- Optional: Code Text: The text box can be used for notes.

It is not needed to complete the following fields:

- Code System Type: This is completed by the system.
- References: It is not necessary to add references.
- URL: This is completed automatically by the system. See Table 6 for the hyperlinks that are used for each code system. (NB. This is not functional in the training instance of EU-SRS.)

Table 6. Overview of identifiers used for proteins

Code system	Code	URL	CV
Common identifiers			
CAS (SCIFINDER)	CAS number e.g. 1610833-03-8	https://chem.nlm.nih.gov/chemidplus/	y
DRUG BANK	Drugbank code e.g. DB14012	https://go.drugbank.com/drugs/[code]	y
EPAR	Product name in small letters e.g. crysvita	https://www.ema.europa.eu/en/medicines/human/EPAR/[code]	y
EVMPD*	EVMPD		y
FDA-UNII	UNII e.g. G9WJT6RD29	https://gsrs.ncats.nih.gov/ginas/app/beta/home	n
INN	INN code e.g. 10301	https://extranet.who.int/soinn/	y
INN RL	INN RL no e.g. 77	https://www.who.int/publications/m/item/inn-rl-[code]	n
KEGG	KEGG code e.g. D10913	https://www.genome.jp/entry/[code]	y
NCI_THESAURUS	NCI Thesaurus code e.g. C119744	https://ncit.nci.nih.gov/ncitbrowser/ConceptReport.jsp?dictionary=NCI_Thesaurus&ns=ncit&code=[code]	y
RMS*	Substance type: - Protein: 200000005020 Domain: Human use: 100000000012, Veterinary use: 100000000013		n
SMSID*	SMSID		n
Other identifiers			
EDQM (KNOWLEDGE BASE)	Monograph number e.g. 0721	https://extranet.edqm.eu/4DLink1/4DCGI/Web_View/mono/[code]	y
INN PL	INN PL no e.g. 115	https://www.who.int/publications/m/item/inn-pl-[code]	y
PUBCHEM	Pubchem code e.g. P0DPI1	https://pubchem.ncbi.nlm.nih.gov/protein/[code]	y
SVGID	SVGID (generated by EUSRS V3.x)		-
UNIPROT	Uniprot code e.g. K4LN57	https://www.uniprot.org/uniprotkb/[code]/entry	y

** SMSID, EVMPD and RMS codes are identifiers from SMS, the EMA database. They are publicly available at https://eutct.eudra.org/ and https://spor.ema.europa.eu/smswi/#.*

4.11 Relationships

Relationships can only be generated for existing data sets.

It is possible to add the target(s) of the protein in the section Relationships. The type of the relationship (e.g. Target -> Inhibitor) should be described and references can be added. Currently relationships are not in scope, but they will be considered in the future. A full list of possible relationships in EU-SRS will be included in the general EU-SRS user guide.

4.12 Notes

Notes are generated automatically and includes system validation messages.

4.13 Properties

The Properties section should be used for:

- Molecular weight of subunits and/or the complete protein.
- Molecular formula of the different subunits and/or whole protein.

For each property the Name, the PropertyType, the Amount, the Referenced Substance and one or more References should be filled in. Optional fields are Parameters, whether a property is Defining and options for Access (protected, SMS or admin).

Information about molecular weight is mentioned in the confidential dossier or in public sources. The system automatically adds a calculated molecular formula and molecular weight based on the amino acid sequence in the Subunits section.

See Table 7 for an overview of common properties of proteins.

Table 7. Overview of common properties of proteins

Name	PropertyType	Amount
MOL_WEIGHT: NUMBER(CALCULATED)	CHEMICAL	ESTIMATED 145000 Da (average)
Mol_weight: LC reduced	CHEMICAL	EXACT 234250 Da (average)
Mol_weight: HC deglycosylated reduced	CHEMICAL	EXACT 49018 Da (average)
Mol_weight: intact deglycosylated	CHEMICAL	EXACT 144863 Da (average)
Molecular formula	CHEMICAL	C6452H9958N1722O2010S42 without posttranslational glycosylation, but including N-terminal pyroglutamate formation and lysine residues at the C-terminals of the heavy chains

4.14 References

References can be included directly during the registration of names, codes or relationships, etc. Alternatively, registration can be started by recording all references, which are then used in the course of registering further information. References can be added, updated, or removed from this card.

See List of relevant databases for substances for a list of relevant databases for substances. See Table 8 for an overview of common references of proteins.

For each reference the following fields should be completed:

- Source Type: See Table 8 for an overview of references used for proteins.
- Source Text/Citation: See Table 8 for the correct description. This is usually the name of the source.
- Public Domain: Public Domain checkbox should be added for all public references.
  All public databases are public domain. European Procedure Number is also public domain; the dossier is confident, but the dossier number is public.
  Every public name must have at least one reference with Public Domain checkbox.
- Access: Choose one of the options ('PROTECTED', 'SMS' or 'Admin') when it is not a public source (NB. Only when the check box of Public Domain is empty).
- Tags: One reference that is used as definitional reference must get the thag 'PUBLIC DOMAIN RELEASE' for technical reasons. The reference EMA LIST gets this tag and must be added as definitional reference. (See 4.4.5 Names of proteins)

It is not needed to complete the following fields:

- Source Class: This is completed automatically by the system.
- URL: The URL's in the references section can't be maintained automatically. So use only URL's when really needed for example for articles, but not for public databases.
- Source Id: stays empty. The Source Id is not shown in the record and not used by the system.
- Upload a Document: This option is currently not in use.

Table 8. Overview of references used for proteins

Source type	Citation	Comment	CV
Common references			
CAS	CAS SciFinder	Reference for systematic name https://scifinder-n.cas.org/	Y
CHEMID	ChemIDplus	https://chem.nlm.nih.gov/chemidplus/	Y
DRUGBANK	DrugBank	https://go.drugbank.com/	Y
EMA LIST	SMS	Reference for SMS name. (just for public names) https://spor.ema.europa.eu/smswi/#	Y
EPAR	EPAR	Reference for brand name https://www.ema.europa.eu/en/medicines/human/EPAR/	N
European Procedure Number	Procedure number e.g. EMEA/H/C/001109	Definitional reference. No link.	Y
FDA_SRS	FDA GSRS	https://precision.fda.gov/uniisearch	Y
INN List	INN recommended List No.	Reference for INN name. The number of the published recommended INN list. https://extranet.who.int/soinn/	Y
INN List	INN description	Reference for INN description. https://extranet.who.int/soinn/	

Source type	Citation	Comment	CV
KEGG	KEGG	https://www.genome.jp/kegg/	Y
NCI THESAURUS	NCI Thesaurus	https://ncithesaurus.nci.nih.gov/ncitbrowser/	Y
Other references			
AAN	AAN	Australian Approved Name	Y
BAN	BAN	British Approved Name	Y
BRITISH PHARMACOPOEIA	BP year	e.g. BP2022	Y
EMA List	SVG	Reference for public names without valid public source. https://spor.ema.europa.eu/smswi/#	Y
EUROPEAN PHARMACOPOEIA	Ph. Eur.: Issue, monograph number	e.g. Ph. Eur.: 10.0, 2118 https://pheur.edqm.eu/app/10-0/search/	Y
G-SRS	G-SRS (NCATS)	https://gsrs.ncats.nih.gov/ginas/app/beta/browse-substance	N
INCI	INCI	https://ec.europa.eu/growth/tools-databases/cosing/	N
INN List	INN proposed List No.	Reference for pINN name. The number of the published proposed INN list. https://extranet.who.int/soinn/	Y
JAN	JAN	Japanese Accepted Names https://jpdb.nihs.go.jp/jan/index.aspx	Y
MARTINDALE	Martindale year	e.g. Martindale 2022 https://www.new.medicinescomplete.com/#/	Y
NCI DRUG DICTIONARY	NCI Drug Dictionary	https://www.cancer.gov/publications/dictionaries/cancer-drug	Y
ORPHAN DRUG	Orphan.desig:FDA Orphan.desig:EU		Y
PUBCHEM	PubChem	https://pubchem.ncbi.nlm.nih.gov/	Y
STN (SCIFINDER)*	STN	Reference for CAS ID in code section.	Y
UNIPROT	UniProt	https://www.uniprot.org/	Y
USANCOUN	USAN	To be used for the USAN designation. https://searchusan.ama-assn.org/finder/usan/search/*/relevant/1/	Y
USANCOUN	USAN description	To be used for USAN chemical names. https://searchusan.ama-assn.org/finder/usan/search/*/relevant/1/	Y
USP/NF	USP issue	e.g. USP42 https://online.uspnf.com/uspnf	Y

** STN (SCIFINDER) is automatically added as reference of CAS.*

4.15 Completion of Registration

To complete the registration of "Protein", a "Definitional Reference" must be inserted. If the record is allowed to be published, the Access lock icon should be opened.

The data set is not saved and inserted into the database until "Submit" is selected. A dialogue window appears (Figure 13). If all messages are highlighted in green, click on "Dismiss All" and then on "Submit". The data set can now be viewed with "View Substance" in the browse mode.

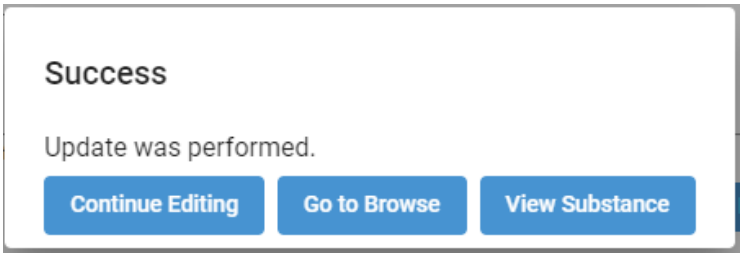


Figure 13. Notification of completed substance registration.

If messages with a yellow background appear, these details must first be checked for correctness. Registration can still take place. Messages with a red background do not allow registration. These errors must first be corrected in the data set so that registration can subsequently take place.

NOTE: Please make sure that the editing of the substance mode does not exceed 120 min. If this time is exceeded, submission is not possible.

4.16 Verification of completed record

After completion of the registration, the whole record should be verified in browse mode (Figure 14). If the record has been verified, it should be approved and pushed to the production environment.

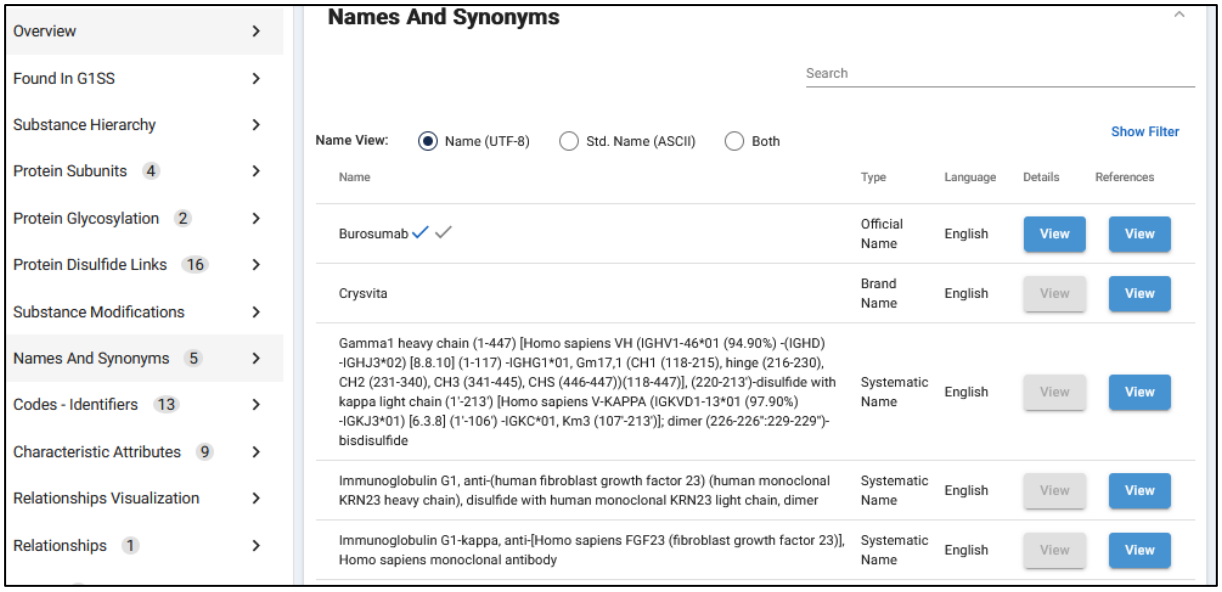


Figure 14. Extract from the browse mode.

4.17 Change Reason

If a protein record is updated, the reason for creating a new version must be recorded under Change reason. This tab appears only from the second version. So, it is not present at the first registration. In order to simplify the future search, a short valid reason should be given if possible. Examples of change reasons are summarized in Table 9. Selecting multiple change reasons is possible, e.g. name, codes, relationships.

Table 9. Examples of change reasons.

Change reason	Comment
Code	Code was added/removed
Disulfide links	Disulfide links were changed
Modification	Modification was added/removed
Name	Name was added/removed
Properties	Properties were added/removed
PT	Preferred Term was changed
Sequence	Amino acid sequence was changed
Substance	Whole review of the record
References	References were added/removed
Relationship	Relationship was added/removed

5 Building SSG1 records with glycosylation details

Glycosylation details are captured at Specified Substance Group 1 (SSG1) level.

5.1 Duplicate check

The first step of any registration should be to ensure that the substance is not already in EU-SRS**,** to prevent from adding a duplicate. To ensure there are no duplicates, different elements can be used for verification, such as names, identifiers and hierarchy of protein. The approach for duplicates check should be determined case-by-case. If duplicates are found, further investigation is required to determine if it's a true duplicate.
There are several ways to search for protein substances in EU-SRS: Global Search, Advanced Search and Sequence Search.

5.2 General information

Each substance is unique in EU-SRS. Select "G1 Specified Substance" from the "Register" menu. The SSG1 registration form will be displayed (see Figure 15). Section cards are collapsible to ease navigation, however the elements in this form are:

- Overview
- Names
- Constituents
- Agent Modifications
- Structural Modifications
- Physical Modifications
- Codes
- Relationships
- Notes
- Properties
- References

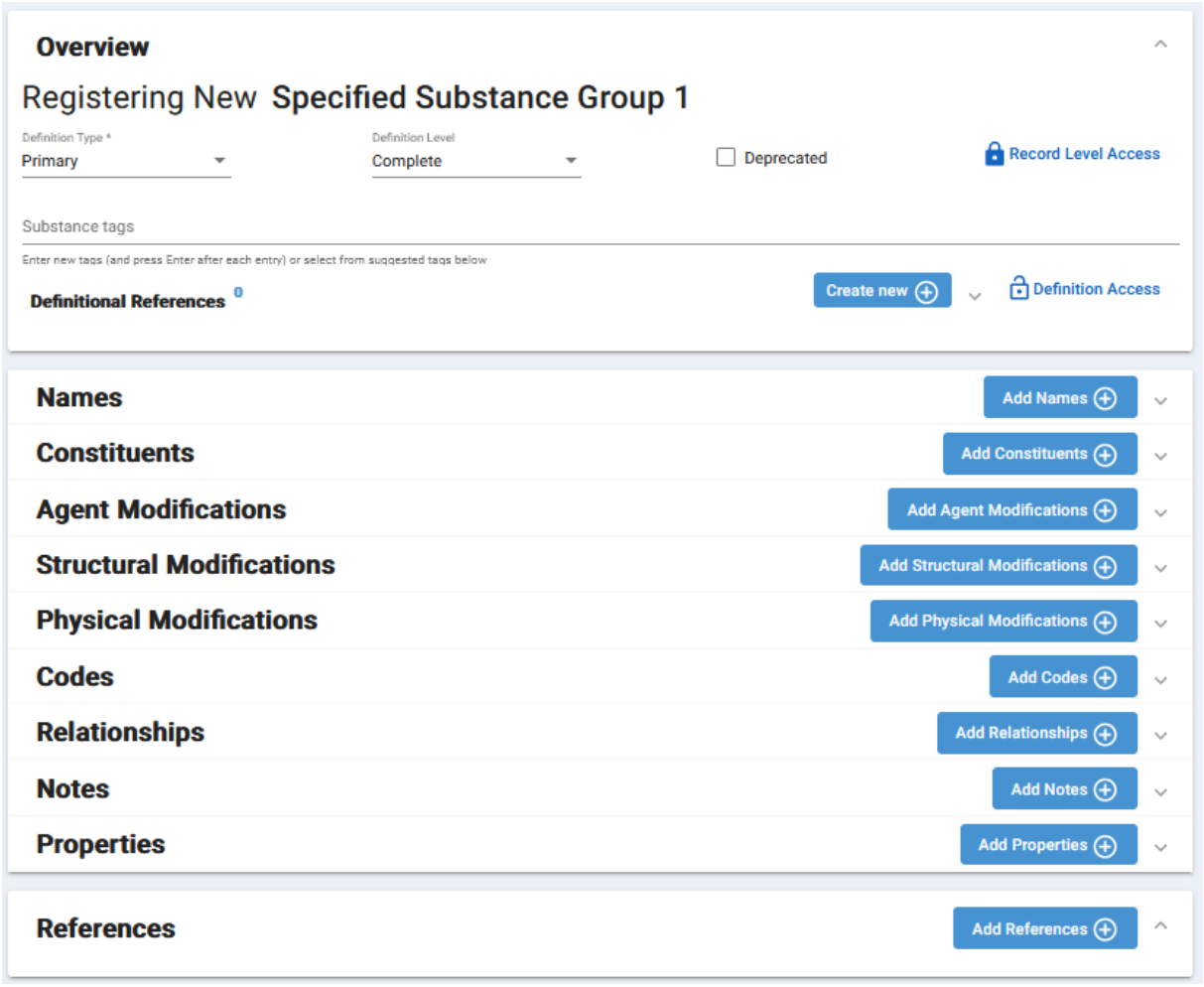


Figure 15. A form for registering a new Chemical in EU-SRS.

Mandatory fields for SSG1 records are:

- at least one name with an unprotected reference marked “Public Domain” in order to be made public.
- at least one definitional reference
- at least one constituent; this is the related protein record

5.3 Overview

See 4.3.1 for an explanation of the data fields.

When registering an SSG1 record complete the Overview section:

- Definition type is ‘Primary’.
- Definition level is ‘Complete’ when the complete amino acid sequence is known. The amino acid sequence is captured in the protein record.
- Definitional Reference. For technical reasons at least 1 reference with the tag ‘PUBLIC DOMAIN RELEASE’ must be added as definitional reference. This is the reference ‘INN List’ with citation ‘INN recommended List No. xx’). Another definitional references for SSG1 records of proteins is ‘European Procedure Number’.

5.4 Names

See 4.4 for an explanation of the data fields.

Names must be written in small letters with a capital at the beginning. Greek letters, Roman numerals, superscript and subscript must be written in HTML meta-language (see HTML code elements for displaying special characters). At least 1 public reference must be added to each name. The check box 'Display Name' is used for the preferred name in EU-SRS.

The preferred name must be written as 'INN-suffix' (e.g. 'Anifrolumab-fnia'). The suffix is the 4-letter suffix of random letters given to biologic medication by the FDA. This suffix helps to differentiate between biosimilars. When there is no 4-letter suffix available, the sequential code INN-EUxxx' will be given ('EU001' etc). The type is 'Common Name'.

5.5 Constituents

The SSG1 record is used for glycosylation details. The related protein record with the defining elements of the protein (amino acid sequence, disulfide links, site and type of glycosylation and essential modifications) must be added as a constituent to get the correct substance hierarchy in EU-SRS. The role added must be Parent substance. The amount can stay empty.

5.6 Structural modifications

The glycan occupation and the specific glycans are added to the structural modifications section.

All glycans with an amount > 1% must be added and all high mannose glycans and glycans with sialic acid.

For each modification fill out the Modification Type, Fragment (after change), Location Type, Sites, Extent ('Partial' or 'Complete') and Extent Amount. For the N-glycan formation the fragment is the glycan and the Extent amount is the amount of that specific glycan.

Glycans in EU-SRS are named according to the Oxford notation system (e.g. A1G1S1F). They get also 'Glycan' in the name and the code from the glytoucan database.

See Table 10 for an overview of common modifications used to describe glycosylation of proteins.

Table 10. Overview of common modifications used to describe glycosylation of proteins

Modification	Modification Type	Fragment	Location Type
N-Glycosylation	N-Glycan occupation	Asparagine	Site-specific
N-Glycosylation	N-Glycan formation	Glycan-G80858MF-A2F	Site-specific
N-Glycosylation	N-Glycan formation	Glycan-G27919IH-A2G1Fa	Site-specific
N-Glycosylation	N-Glycan formation	Glycan-G58667NI-A2G1Fb	Site-specific
N-Glycosylation	N-Glycan formation	Glycan-G88069QW-A2G2F [Gal(b1-4)GlcNAc]	Site-specific
N-Glycosylation	N-Glycan formation	Glycan-G55220VL-M5	Site-specific

5.7 Identifiers

See 4.10 for an explanation of the data fields.

The identifiers section is used for codes.

This section can also be used for references, because it is possible to add hyperlinks that are automatically updated when the link changes. (NB. This is not functional in the training instance of EU-SRS.) The hyperlinks in the reference section can't be updated automatically.

It is also possible to add notes to the codes section. Notes in the identifiers section are more readable than notes in the notes section.

In the edit mode this section is called 'Codes'.

For each identifier the following fields need to be completed:

- Code System: See Table 11 for an overview of common identifiers used for SSG1 records of proteins.
- Type: usually 'primary'.
- Code: See Table 11 for the codes that belong to each code system. The code is shown in the record and also used to automatically create hyperlinks. (NB. This is not functional in the training instance of EU-SRS.)
- Optional: Code Text: The text box can be used for notes.

The following fields do not have to be completed:

- Code System Type: This is completed by the system.
- References: It is not necessary to add references.
- URL: This is completed automatically by the system. See Table 11 for the hyperlinks that are used for each code system. (NB. This is not functional in the training instance of EU-SRS.)

Table 11. Common identifiers used in the SSG1 record of proteins

Code system	Code	URL	CV
Common identifiers			
EPAR	Product name in small letters e.g. crysvita	https://www.ema.europa.eu/en/medicines/human/EPAR/[code]	y
INN	INN code e.g. 10301	https://extranet.who.int/soinn/	y
INN RL	INN RL no e.g. 77	https://www.who.int/publications/m/item/inn-rl-[code]	n
Other identifiers			
INN PL	INN PL no e.g. 115	https://www.who.int/publications/m/item/inn-pl-[code]	y
SVGID	SVGID (generated by EUSRS V3.x)		-
UNIPROT	Uniprot code e.g. K4LN57	https://www.uniprot.org/uniprotkb/[code]/entry	y

5.8 Properties

The following properties should be added as percentage if they are known:

- Total fucosylation
- Total galactosylation
- Total high mannose
- Total sialylation

For each property the Name, the Property Type, the Amount and one or more References should be filled in. Optional fields are Parameters, a Referenced Substance, whether a property is Defining and options for Access (protected, SMS or admin). The Property Type of the properties mentioned above is 'protein' for all of them.

5.9 References

All used references should be added to the section References. There should be at least 1 public reference for the name.

For each reference the following fields need to be completed:

- Reference Type: See Table 12 for an overview of references used for proteins.
- Source Text/Citation: See Table 12 for the correct description. This is usually the name of the source.
- Public Domain: check the box if it is a public source.
  All public databases are public domain. European Procedure Number is also public domain; the dossier is confident, but the dossier number is public.
- Access: Choose one of the options ('PROTECTED', 'SMS' or 'Admin') when it is not a public source (NB. Only when the check box of Public Domain is empty).
- Tags: One reference that is used as definitional reference must get the thag 'PUBLIC DOMAIN RELEASE' for technical reasons. The reference INN List with citation 'INN recommended List No. xx') gets this tag and must be added as definitional reference.

It is not needed to complete the following fields:

- Source Class: This is completed automatically by the system.
- URL: The URL's in the references section can't be maintained automatically. So use only URL's when really needed for example for articles, but not for public databases.
- Source Id: stays empty. The Source Id is not shown in the record and not used by the system.
- Upload a Document: This option is currently not in use.

Table 12. Overview of references used for the SSG1 record of proteins

Source type	Citation	Comment	CV
Common references			
European Procedure Number	Procedure number e.g. EMEA/H/C/001109	Definitional reference. No link.	Y
INN List	INN recommended List No.	Reference for INN name. The number of the published recommended INN list. https://extranet.who.int/soinn/	Y
INN List	INN description	Reference for INN description. https://extranet.who.int/soinn/	Y

5.10 Completion of Registration

To complete the registration of " Specified Substance Group 1", a "Definitional Reference" must be inserted. If the record is allowed to be published, the Access lock icon should be opened.

The data set is not saved and inserted into the database until "Submit" is selected. A dialogue window appears (Figure 16). If all messages are highlighted in green, click on "Dismiss All" and then on "Submit". The data set can now be viewed with "View Substance" in the browse mode.

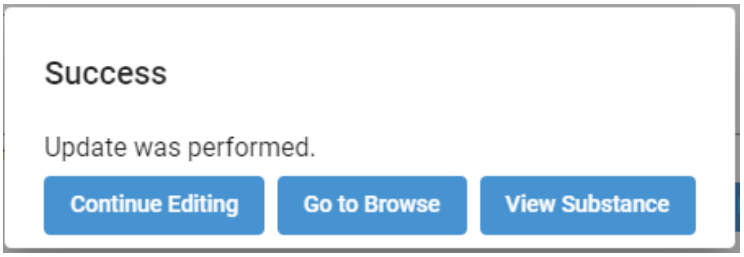


Figure 16. Notification of completed substance registration.

If messages with a yellow background appear, these details must first be checked for correctness. Registration can still take place. Messages with a red background do not allow registration. These errors must first be corrected in the data set so that registration can subsequently take place.

NOTE: Please make sure that the editing of the substance mode does not exceed 120 min. If this time is exceeded, submission is not possible.

5.11 Verification of completed record

After completion of the registration, the whole record should be verified in browse mode (Figure 17). If the record has been verified, it should be approved and pushed to the production environment.

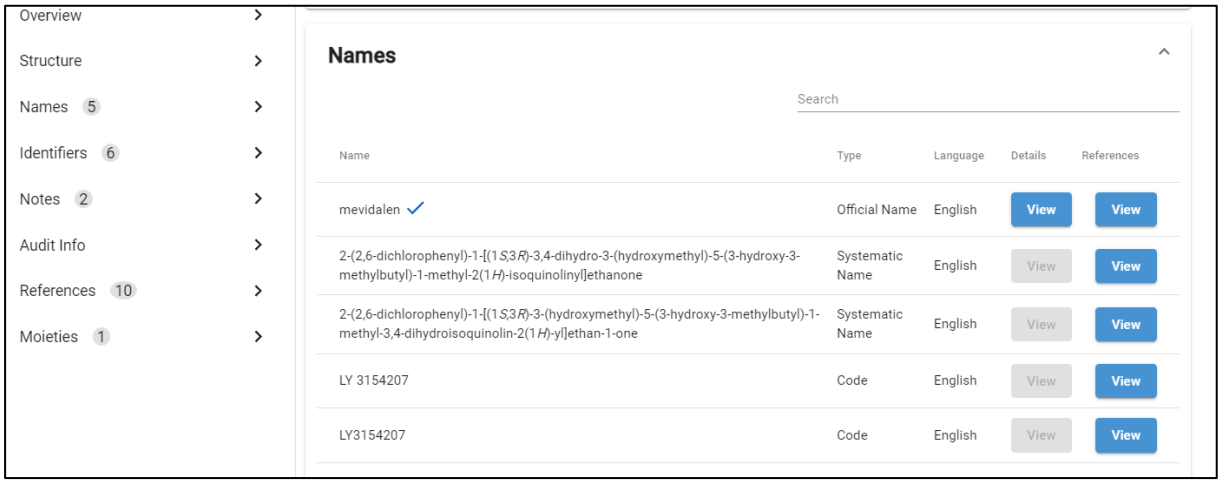


Figure 17. Extract from the browse mode.

5.12 Change Reason

If a substance is updated, the reason for creating a new version must be recorded under Change reason. This tab appears only from the second version. So, it is not present at the first registration. In order to simplify the future search, a short valid reason should be given if possible. Examples of change reasons are summarized in Table 13. Selecting multiple change reasons is possible, e.g. name, codes, relationships.

Table 13. Examples of change reasons.

Change reason	Comment
Name	Name was added/removed
Code	Code was added/removed
Constituent	Constituent was changed
PT	Preferred Term was changed
Modifications	Modifications were added/removed
Substance	Whole review of the record
References	References were added/removed
Relationships	Relationship was added/removed
Properties	Properties was added/removed

6 Protein Types

This chapter summarizes specific properties of the various protein types that are relevant when building protein records and SSG1 records with glycosylation details in EU-SRS. This chapter contains at the moment only information about monoclonal antibodies, but will be expanded in the future.

6.1 Monoclonal antibodies

Subunits: In general subunit 1 and 2 are two identical heavy chains and subunit 3 and 4 are two identical light chains. In case of bispecific antibodies the heavy and light chains are not identical.

Glycosylation: Typical for monoclonal antibodies is Glycosylation Type 'Mammalian' and an N-glycosylation site on the heavy chains. Add the N-glycosylation sites in all subunits even if the subunits are identical (in case of identical heavy chains and/or light chains).

Information about glycan distribution can be found in the recommended INN-list and in the confidential dossier.

7 Appendix

7.1 HTML code elements for displaying special characters

HTML Code	HTML View
₂	x_2
²	x^2
←	←
→	→
<	<
>	>
±	±
Α	Α
α	α
Β	Β
β	β
Γ	Γ
γ	γ
Δ	Δ
δ	δ
Ε	Ε
ε	ε
Ζ	Ζ
ζ	ζ
Η	Η
η	η
Θ	Θ
θ	θ

HTML Code	HTML View
Ι	Ι
ι	ι
Κ	Κ
κ	κ
Λ	Λ
λ	λ
Μ	Μ
μ	μ
Ν	Ν
ν	ν
Ξ	Ξ
ξ	ξ
Ο	Ο
ο	ο
Π	Π
π	π
Ρ	Ρ
ρ	ρ
Σ	Σ
σ	σ
Τ	Τ
τ	τ
Υ	Υ
υ	υ
Φ	Φ
φ	φ

HTML Code	HTML View
Χ	Χ
χ	χ
Ψ	Ψ
ψ	ψ
Ω	Ω
ω	ω
±	±
ß	ß
Ä	Ä
ä	ä
Ö	Ö
ö	ö
Ü	Ü
ü	ü
§	§
¯	¯
Ⅰ	Ⅰ
Ⅱ	Ⅱ
Ⅲ	Ⅲ
Ⅳ	Ⅳ
Ⅴ	Ⅴ
Ⅵ	Ⅵ
Ⅶ	Ⅶ
Ⅷ	Ⅷ
Ⅸ	Ⅸ
Ⅹ	Ⅹ

HTML Code	HTML View
&#8554;	Ⅺ
&#8555;	Ⅻ
ⅩⅢ	XⅢ
ⅩⅣ	XⅣ
ⅩⅤ	XⅤ
ⅩⅥ	XⅥ
ⅩⅦ	XⅦ
ⅩⅧ	XⅧ
ⅩⅨ	XⅨ
ⅩⅩ	XⅩ

7.2 Valid source types and citations

Code system	Code	URL	CV
CAS (SCIFINDER)	CAS number e.g. 1610833-03-8	https://chem.nlm.nih.gov/chemidplus/	y
DRUG BANK	Drugbank code e.g. DB14012	https://go.drugbank.com/drugs/[code]	y
EPAR	Product name in small letters e.g. crysvita	https://www.ema.europa.eu/en/medicines/human/EPAR/[code]	y
EVMPD*	EVMPD		y
FDA-UNII	UNII e.g. G9WJT6RD29	https://gsrs.ncats.nih.gov/ginas/app/beta/home	n
INN	INN code e.g. 10301	https://extranet.who.int/soinn/	y
INN RL	INN RL no e.g. 77	https://www.who.int/publications/m/item/inn-rl-[code]	n
KEGG	KEGG code e.g. D10913	https://www.genome.jp/entry/[code]	y
NCI_THESAURUS	NCI Thesaurus code e.g. C119744	https://ncit.nci.nih.gov/ncitbrowser/ConceptReport.jsp?dictionary=NCI_Thesaurus&ns=ncit&code=[code]	y
EDQM (KNOWLEDGE BASE)	Monograph number e.g. 0721	https://extranet.edqm.eu/4DLink1/4DCGI/Web_View/mono/[code]	y
INN PL	INN PL no e.g. 115	https://www.who.int/publications/m/item/inn-pl-[code]	y
PUBCHEM	Pubchem code e.g. P0DPI1	https://pubchem.ncbi.nlm.nih.gov/protein/[code]	y
SVGID	SVGID (generated by EUSRS V3.x)		-
UNIPROT	Uniprot code e.g. K4LN57	https://www.uniprot.org/uniprotkb/[code]/entry	y

7.3 List of relevant databases for substances

When additional information concerning a substance is needed, the following databases can be used as a reference:

Database	Description
INN	The INN Programme assigns International Nonproprietary Names to medicinal substances through a broad consultative process. WHO is responsible for the INNs.
European Pharmacopoeia	The purpose of the European Pharmacopoeia is to promote public health by the provision of recognised common standards for the quality of medicines and their components. As these standards ensure that medicines reaching the market are safe for use by patients, it is essential that they are appropriate. Their existence also facilitates the free movement of medicinal products in Europe and beyond.
Medicines Complete	A site which guides on to several different publications, databases containing information about medicines.
Inxight: Drugs	Site provided by NIH, National Center for Advancing Translational Sciences. Information about e.g. treatment and pharmacology.
FDA Substance Registration System	Registration system in the U.S. by FDA and the U.S. National Library of Medicine (NIH), provides UNII-codes, Unique Ingredient Identifier.
G-SRS	Database built by GiNAS, NIH. This is the basis for the EU-SRS.
United States Approved Names	This is a site for USAN, where to find the approved names, provided by American Medical Association, AMA.
Japanese Accepted Names	This is a site for JAN, where to find the approved names, as part of the Japanese Pharmacopoeia.
PubChem	Chemical information from authoritative sources provided by U.S. National Library of medicine, NIH
European Union Food Additives	This database can serve as a tool to inform about the food additives approved for use in food in the EU and their conditions of use. It is based on the Union list of food.
EU CosIng	CosIng is the European Commission database for information on cosmetic substances and ingredients.
European Chemicals Agency	ECHA is an agency of the European Union and the site provides data from registration dossiers.
EC Active substance database	Site from the European Commission and it provides General index of products by active substance.
Merck Index	Online version of the Merck index, regarded as the most authoritative and reliable source of information on chemicals, drugs and biologicals. Now this trusted resource is available online from the Royal Society of Chemistry.
EU Orphan Database	Site from the European Commission and it provides The Community Register of orphan medicinal products.
FDA Orphan substance database	Site from FDA and it provides The Community Register of orphan medicinal products.
Index Nominum	This is an International Database of Pharmaceutical Substances and Preparations, provided by Wissenschaftliche Verlagsgesellschaft Stuttgart

Database	Description
International Pharmacopoeia	International Pharmacopoeia provided by WHO.
Scifinder	Research discovery application that provides integrated access to the world's most comprehensive and authoritative source of references, substances and reactions in chemistry and related sciences.
FDA Inactive Database	Site provided by FDA with a database of Inactive ingredients.
Uniprot	high-quality, comprehensive and freely accessible resource of protein sequence and functional information
KEGG	Kyoto Encyclopedia of Genes and Genomes
NCI drug dictionary	Drug dictionary with definitions and other information

7.4 EU-SRS field guidance for registering and/or editing a protein

EU-SRS field name	Field details
Overview	
Preferred Term	Decide if the name complies to the PT rules listed in section xx for registration in EU-SRS.
Definition Type	Always Primary.
Definition Level	Complete when the complete amino acid sequence is known
Deprecated	Relevant for a Deprecated Record.
Record Level Access	If no box is ticked the record is public. If it should be non-public choose 'PROTECTED'. 'PROTECTED' means that the record cannot be exported.
Substance tags	*Not applicable.*
Definitional References	Field linked to Reference section.
Names	
Name (Display Name/Preferred Term in EU-SRS)	Each record needs a Preferred term, which is the displayed name in EU-SRS. Necessary details: Choose a name type, Language always 'English', tick 'DN' (= Display Name)', choose a naming organisation, check if a public reference is given. Detailed information on the single fields can be found below.
Name (Additional name in EU-SRS, could be PT in SMS)	May be added, but the Display Name is sufficient. If there is a record in SMS (which is not necessarily the case), the common name used as PT in SMS should always be included as alias in the record as an Additional Listing Name ('Common Name').
Type	Each name type can be chosen, as appropriate. If it is possible to choose between more than one name type, the following priorisation should be used: Official Name, Common Name, Systematic Name, Company Code, Brand Name. The Official Name is any name used by an Official Naming body. An Official Naming Body is any organization allowed to Name a Substance. PT = appropriate name type (e.g., 'Official Name', 'Systematic Name'), Alias = PT in SMS = 'Common Name', Other Alias = 'Common Name', 'Systematic name' or 'Brand name
Access	If the information in the record is confidential, Access can be set to 'PROTECTED', but the name should have a public and a confidential reference.
DN	DN = Display Name: Tick box to indicate what name should be the Display Name in EU-SRS (= Preferred Term in EU-SRS). Mandatory for the EU-SRS-PT.
AL	AL = Additional Listing Name: Field is mandatory for the SMS-PT (not to be ticked for other aliases). If EU-SRS-PT is the same as SMS-PT both check boxes should be used.
Standardized Name	*Not applicable.*
Languages	Always English.
Domains	*Not applicable.*
Jurisdiction	Field is optional when an 'Official Name' type is selected. Example: USAN would have Jurisdiction United States.
References	At least one public reference is needed per name. References can be newly created or reused.
Naming Organization	Field is only displayed if 'Official Name' type is chosen. If displayed it is conditional. Mandatory when the organization are INN, EP., USAN, BAN, JAN, AAN, INCI.

EU-SRS field name	Field details
Protein details	
Protein type	See chapter 2.2 for the protein types
Protein sub type	See chapter 2.3 for the protein sub types. Note: It is possible to select multiple protein subtypes.
Sequence origin	Is for example ‘human’
Sequence type	Is ‘complete’ if the full sequence is known.
Subunits	
Add subunits	Add the amino acid sequences of all subunits. The agreement is to start with the longest subunit and to end with the shortest subunit.
Disulfide links	
Add disulfide links	All existing intra- and interchain disulfide bonds should be added.
Glycosylation	
Glycosylation type	Add the Glycosylation Type (e.g. mammalian) and the sites of N-, C- and O-glycosylation. Do this by selecting a site from the sequence.
Structural modifcations	
Molecular fragment	If only one fragment is formed, add the formed fragment (e.g. pidolic acid is the fragment of N-terminal pyroglutamate formation). If multiple fragments are formed or if the formed fragment is unknown, add the modified residue as fragment (e.g. asparagine as fragment of asparagine deamidation). See Table 2 for an overview of common structural modifications with their fragments.
Modification type	Choose the modification type, see Table 3 for an overview of common structural modifications.
Extent	Choose between ‘partial’ and ‘complete’.
Amount	If the extent is partial fill out the amount.
Location	The options are ‘C-TERMINUS’, ‘N-TERMINUS’, ‘3’-TERMINUS’, ‘5’-TERMINUS’, ‘SITE SPECIFIC’, ‘RESIDUE SPECIFIC’ AND ‘UNKNOWN’. Note: Select ‘Site specific’ when the exact location is known and select ‘Residue specific’ if only the residue is known and not the exact location.
Sites	When the location is known, choose the correct site(s) in the sequence(s).
Residue modified	If location type is ‘RESIDUE SPECIFIC’, add the modified residue.
Access	If the information is confidential, Access can be set to ‘PROTECTED’.
Codes	
Code system	SMSID is always mandatory unless an SMSID is not available - then the SMS team will be asked for. Preferably other public sources are added when available, e.g., INN, CAS, , Wikipedia. Note: See table 3 for full list of common code sources. Classification: Besides the codes from the source databases used, the classification of the substance is entered under code section, which is relevant for searching. To enter the ‘category’ or ‘domain’, RMS is selected from the drop-down list.
Code System Type	*This is a default value (automatically filled in and managed by Admin).*
Type	In nearly all cases default 'Primary' (code uniquely describes the substance). When the code is tied to a group/used for classification (Code system ‘RMS’) use ‘Generic (Family)’. Superseded is set if the code is already deprecated.
Code	Alpha numeric value. Classification: Relevant is the code for the domain ‘Human use’ or 'Veterinary use' and the category ‘Polymer'

EU-SRS field name	Field details
Access	Set to Public from the system after the import from SMS (standard = Public).
URL	Will be generated automatically
Code Text	Optional: for notes.
Comments	*Not applicable.*
References	For CAS ID add STN reference
Relationships	
Related Substance	Can be selected directly by the search from the dataset
Type (purpose hierarchy)	Is selected if the substance is to be assigned to a higher-level data set.
Type (purpose not hierarchy)	All relations that have an arrow in their name can be registered to both basis substance and the related record. Other record gets the reverse relations added automatically.
Access	*Not applicable.*
Mediator Substance	*Not applicable.*
Qualification	For example: Ph. Eur. is registered for Ph. Eur. impurities of monographed substances (for relationship: 'Impurity à Parent').
Interaction Type	Investigation described in the monograph are specified (e.g. HPLC) (for relationship: 'Impurity à Parent').
Notes	
Note	Field is optional and automatically populated by the system.
References	*Not applicable.*
Access	*Not applicable.*
Properties	
Name	The Molecular weight of subunits and/or the complete protein and the Molecular formula of the different subunits and/or whole protein can be added here. Note: see Table 4 for an overview of common properties of proteins.
Property Type	Chemical, Enzymatic or Physical can be selected as Property type from the drop-down list
Defining	Default is empty Tick box. The check mark can be set if the requirements are mandatory.
Referenced Substance	*Not applicable.*
Parameters	*Not applicable.*
Amount	Appears as a pop-up window. The type information 'mol ratio' or 'weight ratio' can be registered, as mean value (Average), or range (Low Limit and High Limit). Units can be selected, but it is not mandatory. Access is set to Public from the system after the import from SMS (standard = Public).
References	*Not applicable.*
Access	*Not applicable.*
References	
Source Type	See Table 8 for an overview of common references of proteins; Mandatory field. Additional value not in CV is possible as temporary value (note: be careful with adding a new value, this needs to be communicated with the technical team).
Source text/Citation	Mandatory field and should represent the related value of the type. See also Table 8 for source text/citation.
Public Domain	Default is Public (Tick box), but it may be set 'Non-Public' in combination with a public reference.
Access	Default is ‘Public’. Confidential = tick ‘PROTECTED’.
URL	This field is optional.
Source Id	*Not applicable*.
Upload a Document	This field is optional.

EU-SRS field name	Field details
Tags	This field should be populated by at least one value (‘Public domain release’) and if applicable other values.
Change Reason	
Change Reason	Add the reason of the change of the record that led to the creation of a new version. Use short notes (if something was added, edited or removed), e.g. Name, Code, RMS, Reference, Naming Organization.