**Global Substance Registration System: Consistent scientific descriptions for substances related to health**

Tyler Peryea1,2, Noel Southall2,\*, Mitch Miller2, Danny Katzel2, Niko Anderson2, Jorge Neyra2, Sarah Stemann2, Dac-Trung Nguyen2, Dammika Amugoda2, Archana Newatia1, Ramez Ghazzaoui1, Elaine Johanson1, Herman Diederik3, Larry Callahan1, Frank Switzer1

**Supplementary Data**

***Substance Data Model***

The information system is designed around the 6 types of substances referenced in the ISO 11238 standard: chemicals, nucleic acids, proteins, polymers, structurally diverse, mixtures. The standard also provides models for further specified substances, which are being developed incrementally, beginning with support for Group 1 (support for multiple substance ingredients and physical form). Future support is planned for Group 2 (limited manufacturing info), Group 3 (grade for material), and Group 4 (detailed manufacturing process specifications/formulation).

Each of these substance types, and all relevant data fields present in the ISO standard and its technical implementation guide[1] are included within the system’s data model. Some substance definitions can be represented by multiple substance types and in these cases, it is possible to provide an alternative definition. For example, triptorelin ethylamide is registered as a chemical (UNII: 48S2W73QPR) and an alternative definition providing a sequence-based protein record is also provided (<https://gsrs.ncats.nih.gov/app/substance/48S2W73QPR#altrelationships>). In addition, the system allows the entry, search, and retrieval of certain additional general substance information (e.g. properties and substance relationships). To do this, the system provides a generic method to link many diverse data types to a given substance, in a meaningful and standard-compliant manner. These relationships include linking substances to other substances including cellular metabolites, manufacturing impurities and putative drug targets as well as to different salt forms used to formulate products.

***Substance Names and Identifiers –*** A global substance identifier is needed because official name use differs between jurisdictions or context of use. For example, acetaminophen (UNII: 362O9ITL9D) in the US is referred to as paracetamol elsewhere. Sometimes the same ingredient has different names depending on what kind of product it is used in. For example, palmitic diethanolamide (UNII: SZ3923Q89F) is listed as palmitamide DEA when used in cosmetic products. GSRS tracks the jurisdiction and context of use for official names. GSRS also captures common names, brand names, systematic names, company codes and other identifiers such as registry numbers. All names and identifiers are referenced.

***Chemicals –*** Molecular structure information for chemicals is captured with a MOL V2000 formatted chemical table file that provides an atom list, bond table and additional structural annotations. The molecular structure part of the record is incomplete for interpreting chemical definitions from GSRS. For example, stereochemistry notation in this context can be ambiguous and insufficient for certain cases (e.g. atropisomerism). Consequently, the chemical depiction business rules developed in previous registration systems have also been applied here[2] in combination with additional defining elements as provided in the ISO 11238 standard. Efforts are made to ensure the chemical structures are accurate and descriptive. The drawings and updates of these structures are typically done by hand by experts, and they may employ many different tools to adjust the coordinates.

***Proteins –*** Protein substances are defined based on the amino acid sequences of their subunits, their glycosylation type and sites, disulfide linkages, and other physical or chemical modifications. Modifications are defined in the ISO 11238 standard as irreversible changes in the underlying structure of polymers, proteins, nucleic acids, structurally diverse material or mixture substances and are represented by modification elements. Modifications may be physical, chemical, enzymatic etc. Modifications are often described by amino acid substitutions to the underlying material when definitive structural modifications occur, but the actual position of substitution may be unknown or variable. Physical treatments and agent modifications that result in irreversible structural modifications are also captured.

***Nucleic acids –*** Nucleic acid substances are defined by the sequence of the nitrogen-containing bases and the specific sugar or other covalent moieties that link them along with any modifications of base structures.

***Polymers –*** Polymer substances are polydisperse materials defined based upon the monomers that were used to synthesize them, the structural subunits that constitute them, the modifications that may have been applied to them, and any defining properties like viscosity or density.

***Structurally diverse –*** Structurally diverse substances are defined based upon the identity of the parent substance, usually its taxonomy, e.g. family, genus, species, and author for organisms or source material class or type, and then further delineated by other definitional properties such as infraspecific type, part, fraction or modification as needed.

***Mixtures –*** Related substances either prepared together or the result of the same synthetic method. Mixture substances are defined based upon their components – the identity of each component substance is separately registered in the system as well as the role of the component within the mixture (roles can be one of: May be present [any of]/May be present [one of]/Must be present [all of]) or the identity of a parent substance.

***Substance concepts and further specified substances*** – In addition to the classes of physical substance ingredients listed above, it can be convenient to use the same data models to capture definitions of substance concepts used in regulation which may refer to aspects of a physical substance or provide an incomplete definition. On the other hand, specified substances provide further definition to a substance, distinguishing for example physical form or context. One example would be the difference between the chemical sodium chloride (UNII: 451W47IQ8X) and the ISO 11238 group 1 specified substance isotonic sodium chloride solution (UNII: VR5Y7PDT5W).

***Relationships and Properties*** ***–*** GSRS captures an extensive number of relationships between substances. These include relationships from a given active substance to its target, metabolic enzymes, transporters, off target interactions, metabolites and impurities. It also captures a variety of physical, biological properties inherent in the substance. Occasionally, these relationships can be defining for a substance.

***Unique Ingredient Identifier –*** The FDA supports health information technology initiatives by generating unique ingredient identifiers (UNIIs) for substances in drugs, biologics, foods, cosmetics, dietary supplements, tobacco products, and devices. The UNII is a non-proprietary, free to use, unique, unambiguous, non-semantic, alphanumeric identifier based on a substance’s defining properties from the ISO 11238 data model. GSRS is the software that generates UNIIs used in FDA electronic listing as seen in DailyMed.[3] It is also used for other regulatory activities throughout product life cycles, encompassing clinical trial phases, product marketing and post-market surveillance. New UNII requests, data issues, or questions can be addressed by contacting FDA-SRS@fda.hhs.gov.

***Software Architecture***

GSRS evolved from initial experience developing a prototype application from an existing project[4]. Software requirements for GSRS included for it to be self-contained and modular, with well-defined data access application programming interfaces (APIs) for content updates and curation and anticipating third-party client access. GSRS requirements also include access control and an audit trail for every piece of information accessed through a web browser. Registration would have to be configurable, to enforce local “business rules” for data presentation and generating new identifiers.

In February 2013, the FDA Office of the Chief Health Scientist (Informatics), and the National Center for Advancing Translational Sciences (NCATS) at the National Institutes of Health (NIH) organized a meeting at U.S. Pharmacopeia headquarters to discuss with other interested groups the creation of such an open-source substance information and registration system. From the meeting it was clear that there is a global need for such a system, and a strong interest by others in participating in the development of such a system.

Several national information systems were described, including regional requirements which any global system would need to support. FDA restated their commitment to develop and maintain a distributable data system that is compliant with ISO 11238 to eventually replace their existing systems. Staff from Netherlands Medicines Evaluation Board (MEB), Germany’s Federal Institute for Drugs and Medical Devices and Health Canada all emphasized that regional knowledge and input on the nomenclature, laws and guidelines would be essential for having any future system be globally useful. Additional groups participating in the discussion included International Union of Pure and Applied Chemistry (IUPAC; authors of InChI[5]), the Royal Society of Chemistry, the National Library of Medicine (PubChem[6] and ChemID*plus*[7]), the National Cancer Institute (including CACTVS[8]), Kew Gardens and the United States Pharmacopeia. All of these organizations expressed support for the project and presented on information that they thought would be of use to our effort. NCATS demonstrated a working prototype application to register chemical substances, but the data model did not yet support the full range of substances anticipated by the ISO 11238 model or satisfy the current and existing requirements of national authorities.

The Software is designed around Object Oriented principals and extensibility. Code modules for searching, rendering and exporting data are pluggable so that custom functions can be coded and added or integrated with other systems of a regulatory agency. The representational state transfer (REST) API allows GSRS stakeholders to customize their own user interface (UI). The complex stereochemical computations performed inside GSRS have been factored out using molwitch[9] so that they don’t rely on any particular cheminformatics library. It currently supports Jchem[10, 11] and CDK[12]. This allows users to use the cheminformatics library that they have a license for or an open source alternative. GSRS application provides a data access protocol for the web based on the Open Data Protocol[13] and REST. Data exchange is done via Javascript Object Notation (JSON) encoded objects. JSON encoding enables clients to read and write data from GSRS in a language and platform agnostic way. For example, there are programs that have been written to interact with GSRS that have been written in Java, Python, Perl, R, even Microsoft Excel Visual Basic Macros.

***Figure 1 abbreviations used*.**

EPA = United States Environmental Protection Agency, ECHA = European Chemicals Agency, EC/EINECS = European Union Chemical Inventory, EVMPD = EudraVigilance Medicinal Product Dictionary, WHO-DD = World Health Organization Drug Dictionary, ChEMBL = Chemical Database of the European Molecular Biology Laboratory / European Bioinformatics Institute, RXCUI = National Library of Medicine’s RxNorm machine-readable code or identifier, NCBI Taxonomy = The National Center for Biotechnology Information taxonomy resource for curated classification and nomenclature, MESH = National Library of Medicine’s Medical Subject Headings, INN = World Health Organization’s International Nonproprietary Names, MPNS = Kew Garden’s Medicinal Plant Naming Service, INCI = International Nomenclature of Cosmetic Ingredients, USAN = United States Adopted Names.

Supplementary Data References

1. International Standards Organization. *ISO/TS 19844: 2018. Implementation guidelines for ISO 11238 for data elements and structures for the unique identification and exchange of regulated information on substances*. Available from: <https://www.iso.org/standard/71965.html>.

2. United States Food and Drug Administration. *Substance Registration System Standard Operating Procedure*. 2007; Available from: <https://www.fda.gov/media/75274/download>.

3. National Library of Medicine. *DailyMed*. Available from: <https://dailymed.nlm.nih.gov/dailymed/>.

4. Howe, E.A., et al., *BioAssay Research Database (BARD): Chemical biology and probe-development enabled by structured metadata and result types.* Nucleic Acids Research, 2015. **43**(D1): p. D1163-D1170.

5. Heller, S., et al., *InChI - the worldwide chemical structure identifier standard.* J Cheminform, 2013. **5**(1): p. 7.

6. Kim, S., et al., *PubChem Substance and Compound databases.* Nucleic Acids Res, 2016. **44**(D1): p. D1202-13.

7. Wexler, P., *The U.S. National Library of Medicine's Toxicology and Environmental Health Information Program.* Toxicology, 2004. **198**(1-3): p. 161-8.

8. Ihlenfeldt, W.D., et al., *Enhanced CACTVS browser of the Open NCI Database.* J Chem Inf Comput Sci, 2002. **42**(1): p. 46-57.

9. NIH/NCATS. *molwitch*. Available from: <https://github.com/ncats/molwitch>

10. ChemAxon. *JChem Base*. Available from: <https://chemaxon.com/products/jchem-engines>.

11. Csizmadia, F., *JChem: Java applets and modules supporting chemical database handling from web browsers.* J Chem Inf Comput Sci, 2000. **40**(2): p. 323-4.

12. Willighagen, E.L., et al., *The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching.* J Cheminform, 2017. **9**(1): p. 33.

13. odata.org. *OData - the best way to REST*. Available from: <https://www.odata.org/>.