OMG Standards Development Organization®

The 2025 OMG Semantic Augmentation Challenge Guidelines and FAQ

The Business Need

Many datasets are published in formats, such as Comma Separated Values (CSV), that provide minimal documentation or schema information, including semantics, beyond a row of column names. This poses a challenge to users who are looking for specific kinds of data, and to recipients who need to integrate them with other data (internal and external) for reliable reporting, AI, compliance, and other uses.

Ontologies and other less structured resources provide a means of expressing data semantics, but most datasets don't reference them.

Hence the need to be able to augment a dataset with metadata that links its elements (columns) to their semantic meaning. As well as being useful documentation, it should be machine-readable enough to drive a transformation of the CSV to a semantic Resource Description Format (RDF) structure for those columns which have an ontology mapping. Which should be repeatable enough to be re-run on a new version of the CSV to create an updated structure, preserving previously allocated identifiers.

Object Management Group's (OMG's) Financial Sector Domain Task Force has members who face this difficulty daily, and so is now launching this Semantic Augmentation Challenge to discover what exists to address this, how usable it is, and to facilitate making such solutions more visible or, if needed, promoting the development of a new standard.

For clarity, the primary objective is a (new or existing) standard metadata format that can be used to augment such datasets. The full scope is not limited to CSV but would include other common formats such as JSON and XML.

The Challenge

For many CSV files, no definitions of meanings are provided, and users are left to guess meaning from column names and data examples. The Challenge asks participants to recommend and demonstrate, using data sources specified below, how best to include references such as context, definitions, and pointers such that the receiver of the file, whether a human or a program, will have the information needed to understand and use it successfully.

For the Challenge we are providing a CSV file consisting of rows of data about FDIC-insured banks in the United States. The file can be downloaded located at this location. To the extent possible, the meaning of the data in each column should be described by a mapping. We anticipate that some columns will be well-described by formal ontologies, while others may be best explained by reference to other resources, such as online dictionary entries. Note that the purpose of this exercise is not to evaluate the quality of any specific mappings suggested, but to evaluate the mapping metadata. In other words, we do not plan to evaluate who chose the "best" mapping for any of these columns. For example, we won't distinguish based on whether "Branch Address" is mapped generally to "location" or more specifically "commercial location". We will consider ergonomics, flexibility, modern tooling, as some but not all the aspects of the file format for mapping.

Participants will produce a file which contains metadata for mapping each column of the dataset, and the dataset rows as a whole, to an externally defined resource. The file should also contain information describing the datasets from a product level, including things such as the dataset provider, platform, method of delivery, product version; such that if the file were found in isolation from a dataset, it might be possible to locate the dataset. Then (requested but not required), use the metadata to transform the CSV data to RDF as compliant with the ontology as is reasonable. We will ask participants to re-execute the transformation on a new version of the dataset (to be supplied later) which uses the same format.

The Motivation

A Prize of US $1,000 will be awarded to the winner.

All submissions will be listed on the OMG Semantic Augmentation Challenge website which will provide an opportunity to showcase submissions and their creators. The shortlist and winner will be highlighted, and submissions featuring distinctive aspects may be given honorable mentions.

Shortlisted submissions will be encouraged to present their solutions live, whether in-person or remotely at the Challenge event and will be able to benefit from feedback (both live and in more detail afterwards).

Submission Details and Required Deliverables

As described above, participants are provided a dataset in CSV format, which they must use. Participants must provide a file that describes what is in **each** of the columns of the CSV. We are also asking for supporting materials about how the format works, and how extensible it would be to other formats beyond CSV.

- We are looking for participants to submit a file that shows a mapping of columns in a machine readable and processable format to common and citable resources. The format used should reflect current best practices, languages, ergonomics, and technologies.
- While not required, we would be especially interested in files that can be described in and/or mapped to RDF.
- The meaning of the columns of the dataset should be mapped to common and citable resources. We will expect at least one mapping to the Financial Industry Business Ontology (FIBO)(https://github.com/edmcouncil/fibo), and to OGC GeoSPARQL (https://opengeospatial.github.io/ogc-geosparql/geosparql11/geo.html) or GeoNames (https://www.geonames.org/ontology/documentation.html), but we also expect citations to a variety of other types of resources that will demonstrate the robustness of the mapping approach, for example:
- Other well-formed ontologies, whether standards or not
- Non well-formed resources that may not be in any kind of ontology, such as wiki pages, online dictionaries, or informal data resources (participants are urged not to use references to proprietary documentation).
- The file should contain a description of the dataset, version, provider (of the dataset) and the details of any cited resources.
- The Challenge is looking for a submission that uses convenient and economical technology(ies), whether novel, already existing, or in combination. Where existing technologies or specifications are used, indicate their identity and provide or link to their details.

- We wish to receive evidence that the mapping file format works, and how broadly it can be used
- Provide the output generated by the format when used with the source dataset.
- Describe any features or limitations with respect to the mapping file.
- Describe the processing environment(s) in which it was run, including versions of software. Comment on how transferable the format is likely to be to other platforms both proprietary and open source.
- Though not required, a video or other media showing the processing would be welcome. An accompanying narrative should describe how it works and the steps that are being taken.
- The submission should include comments on how the format scales with respect to larger datasets.

All IP will remain the IP of the original owner. OMG's intent is to share online, where permission by the owner is granted, certain materials of or about selected submissions, so please indicate what can be used and its IP ownership.

How to Submit

Please fill out this form if you are interested in participating in the challenge:
https://forms.office.com/Pages/ResponsePage.aspx?id=vE-6QwrcaUK1A2TwNjeZ2DenTNcKUnxFmPjaJVeHunVURDNOOTVPVkRaWEtQWk45R0NPMk9CVUtENS4u

Each Submission should consist of a single Zip file. It should be sent to: semanticchallenge@omg.org

The Zip file should include a file called e.g. SubmissionManifest.txt with the following information:

- Primary contact name
- Email

- Phone (optional)
- Other contact names (optional)
- Organization (optional)
- Title for submission (e.g. the name for the metadata format)
- Copyright waiver
- List of files included in the zip and their purpose (unless it's obvious from their names)

The Process

Submissions will be reviewed by a Panel of judges, and a shortlist selected for presentation and discussion at the June OMG Technical Committee meeting at which the winner will be announced. The meeting will be held in Denver, Colorado and online access will also be available.

The panel will be chaired by Richard Robinson, Co-Chair of OMG's Financial Services Domain Task Force and Bloomberg, and Co-Chair of OMG's Financial Services Domain Task Force. Other judges will be announced shortly.

The evaluation criteria are as follows:

- Clarity and usability of the mapping format, especially for non-technical (business) user
- Practical utility and ergonomics of any tool support for creating the mapping, and transforming the data
- Comprehensiveness of documentation, including rationale and benefits
- Discussion of the mapping process and issues and assumptions addressed, especially related to the use of identifiers (URIs) and structures to handle new versions of the data
- Quality of the output RDF

Timeline

- Submission deadline: May 28th
- Shortlist selected: June 2nd
- Revised dataset provided: June 9th (demonstrates that mapping still works)
- Presentations and decision: June 11th 12-1 pm Eastern (physical event will be held in Denver, 10-11 am MT)

FAQ

Please reference the challenge event page for all information: https://www.omg.org/events/2025q2/special-events/omg-semantic-augmentation-challenge.htm

Can I use an existing specification and tooling, even if I don't own it?

Yes, in fact this is welcomed, but please ensure you have the right to use and publicize whatever you submit – either because it's open source or you have permission.

What format of files are acceptable?

Any syntax of RDF is acceptable including Turtle, JSON-LD, N3, N-triples, Quads, Manchester Syntax, and RDF/XML.

Videos should be playable in recent web browsers without the need for a plugin. MP4 is preferred.

What level of documentation is needed?

Whatever you feel would help people understand and evaluate your submission and its benefits. Comprehensiveness is more important than formality, which is why videos are valuable.

If someone from an open source project, or company, submits the software, then should they get the prize on the basis of the software others (also) wrote?

The prize is not for the software but for the submission which includes augmenting the supplied file, describing it and how it was applied, and showing the results. If multiple people collaborated on the submission, then we suggest they split the prize between them on a basis they determine. The prize will be paid out to the lead contact for the submission.

How can I ask further questions?

Send an email to semanticchallenge@omg.org. All questions and responses will be shared, in the interest of all submitters.