Comments on FDTA Proposal

Pete Rivett, Principal, Federated Knowledge LLC
pete.rivett@federatedknowledge.com

I'm an enthusiastic advocate for semantically-rich open data and see the FDTA implementation as potentially a pivotal moment - with the Proposal as a worthwhile first step. The world is seeing the limits of LLM-based AI and waking up to knowledge graphs and semantics.
And there are lots of enthusiastic people to help.
While I'm not a fan of most uses of "Semantic Layer" I believe semantics can be included without the need to rip systems out and start again. However it does entail getting a deep understanding of the original data and not merely papering over what may be a crumbling data stack.

I have spent most of my career in this area, from traditional metadata management and data governance through to the semantic web and AI. I spent 20+ years on the OMG Architecture Board, was a founding Director of the Enterprise Knowledge Graph Foundation (now incorporated into OMG) and was the lead developer of GLEIF's ontology implementation, and a leading light on FIBO. As a data wrangler and implementer of many data transformations I am well aware of the challenges of applying what seems like a great standard.
I have my own knowledge graph consultancy, Federated Knowledge, LLC (FKL) and am currently working with leading companies in the field including: AuditChain Labs (web3 financial disclosure infrastructure building semantics on XBRL), EmergeGen (semantic data out of unstructured documents using sophisticated AI), Semantic Partners (helping businesses with semantic projects), Ethar (making semantic knowledge available via augmented reality), and Adaptive (integrated model management).
I continue to work with OMG on most of the specifications referenced in both this and OMG's own response. However this response represents my own personal opinions and not of any of the above organizations: hence this is more of a *minority report*.

I have several points to make and am going to be terse at the risk of being impolite. I'm always happy to provide more detail on any of these.

Semantics

I applaud the emphasis on "semantics" and "semantic meaning" (which appears 8 times in the Proposal) - however what I find lacking is much in terms of specifics as to what that entails in practice. I think it needs more than some sort of machine-readability or searchability.
I think it would be a huge waste if the result of the FDTA activity were to be definitions of the form "the income of the company ".
As important are precise English definitions and logical definitions or constraints.

A very important point is to separate terms (the words as represented in vocabularies) from concepts (the underlying meaning, as represented in ontologies). Even within the Federal

space, terms have different meanings in different documents (this was amply illustrated by the footnote related to “financial entity” in the Proposal itself), and it’s unrealistic to expect them all to change to a single definition, especially when the terms are defined in laws and regulations. OMG’s **Multiple Vocabulary Facility** specification provides the capability to do this, and associate different Terms with different Communities (which could be different Agencies or even the readers of specific laws).

Given that the same word can have different meanings in different contexts, Agencies should beware of using LLMs trained on widely available documents or web pages from diverse sources.

One specific concern is that the list of acceptable schema formats does not include the ontology languages in universal use by the linked data and ontology communities and managed by W3C, another VCSB. These standards certainly meet the four properties listed and are:

- RDF Schema (RDFS)
- Shapes Constraints Language (SHACL)
- Web Ontology Language (OWL)

While I understand the list is not intended to be exclusive, I think the lack of the above formats will send some people the message that the Agencies are not truly serious about semantics as widely practiced. Indeed FIGI itself is defined as an OWL ontology, and GLEIF publishes both its schema and its LEI data in OWL, based on work I carried out in conjunction with data.world. https://www.gleif.org/en/about-lei/semantic-representation-of-the-lei/lei-model-in-rdf-resource-description-framework
And while I’m pleased to see that “The Agencies also expect to monitor developments related to data standards, including the joint standards, and update the joint rule, as appropriate.” and happy to assist with that, the most recent of the above specifications is over 7 years old.

The Proposal says “Second, data transmission or schema and taxonomy formats that have these properties are likely to be interoperable with each other.” Sadly, I believe that’s easier said than done with the formats listed in the Proposal.
XBRL taxonomies are the subject of section E, though not referenced as such. It’s important to understand that they are very sophisticated (closer in essence to ontologies), need specialized tools to manage (such as Auditchain Labs’ Luca Suite™), and are not at all similar to other taxonomies such as SKOS and the NISO thesaurus specification referenced. And, even though they internally make use of XML Schemas, they’re not interoperable with independent XML Schemas.
Finally, JSON Schema, which seems to be stable and well-run, is managed by a group of experts via their GitHub site: it’s not clear they constitute a VCSB and the formal status and governance seems unclear.
In summary, I think the Agencies should declare specific interoperability formats, subject to regular review, rather than leaving open what could become a free-for-all of different formats claiming compliance with the 4 properties.

Another important point related to semantics is to encourage a “things not strings” approach. For example the *status* of a grant should not be represented by a string such as “Approved” but by a uniquely-identified entity with a precise definition as to what “Approved” means, the source where that meaning is defined, and where it appears in the overall lifecycle of a Grant.

Identifiers

The specifications proposed make sense and the rationale is sound. What is also important is to ensure an open and transparent change management process for the specifications (such as OMG has for any of its specifications) and an error reporting process for the data (such as GLEIF has for legal entity data https://www.gleif.org/en/lei-data/gleif-data-quality-management/challenge-lei-data). While most VCSBs do have this, I can see value in the Agencies putting a spotlight on this aspect.

The power of VCSB-driven change management is illustrated by the fact that OMG has already created versions 1.1 and 1.2 of FIGI to respond to industry comments; and has extended coverage to crypto assets in collaboration with Keiko.

I think it’s important to note that mandating identifiers such as FIGI and LEI does not require expensive replacement of internal processing systems and applications, but a mapping stage that adds the official identifier at the point of submission.
And that nothing precludes an organization using any number of different identifiers (internal or legacy) for the same entity: the linked data specifications were designed for this!

Speaking of mapping, I think the Agencies could usefully apply their clout to induce bodies to make such mappings available as well as the specifications. GLEIF already provides an open mapping between LEIs and ISINs (https://www.gleif.org/en/lei-data/lei-mapping/download-isin-to-lei-relationship-files) and I’d like to see something similar relating FIGIs to the LEI(s) that the instruments apply to. Likewise for datasets made available by stock exchanges such as NASDAQ - which, when I last looked, asked registrants for their LEI but did not provide it in their basic reference data (though FIGI is provided) https://data.nasdaq.com/databases/E360#anchor-reference-data-ndaq-rd- .

Linked Data

To be of true value, I believe the agencies should go beyond identifiers to web-based URIs to allow a common reference point and linking together of data silos. This not only allows the Agencies to interoperate but the community at large to analyze the data.
This is something already available for legal entities from the GLEIF data hosted by data.world already linked to.

Further, OMG has taken the ISO-3166 country codes and created an ontology and set of linked data as part of its Languages, Countries and Codes (LCC) specification

https://www.omg.org/spec/LCC, with Currencies to follow based on existing FIBO work. This has already been utilized in the public LEI dataset previously mentioned: and it allows distributed queries such as discovering those legal entities with headquarters in countries whose administrative language includes French.

Techniques such as Linked Data Fragments https://linkeddatafragments.org/ are available to mitigate the burden on those running servers.

Detailed Comments

The term “hierarchical structure” used several times is ambiguous. It could refer to either:

- a real hierarchical structure in the data e.g. from a legal entity to all its child entities and from those to the stocks they’ve issued
- A taxonomic hierarchy e.g. *FDIC-insured Bank* being a subtype of *Bank* which is a subtype of *Financial Entity*.

Several of the ISO standards referenced do not seem to meet the definition of “open” since ISO charges for the specification documents (although not the data, which is typically managed by another entity). For example the LEI specification, 17442–1:2020, costs $72.83 at today’s exchange rate. ISO does make certain well-used specifications available free of charge https://standards.iso.org/ittf/PubliclyAvailableStandards/index.html and I would urge the Agencies to prevail on ISO to do that for the ones referenced by the Proposal.

The term “data element” is used several times without definition.

Likewise “data asset”, defined as “data sets that may be grouped together” seems to be lacking a motivation or purpose. After all, in general speech, an asset is something of value. I think in practice this would align with “data product’ as widely used in industry and in OMG’s draft Data Products Ontology (DPROD) specification.

In order to allow elements to be ;linked to the legislation that requires them, a laudable goal, the legislation itself will need to be linked, potentially down to the clause level, in a machine-readable way. In order to allow impact analysis of the clause changing.

The Proposal references the need for “verification of data assets”. While it is quite possible to check that machine-readable data meets structural and semantic constraints, and consistency with other sources, it is another matter to verify it against reality. The best that can often be achieved is attestation from responsible individuals, with accountability.

With respect to dates the Proposal states “While date and time information may be displayed on forms, web pages, user interfaces, and other media in other formats (e.g., Month, Day, Year), the underlying machine-readable data should, to the extent feasible, follow the ISO 8601 format.” I think this is over-reach if by “underlying” is meant the data as stored in internal databases. In many cases, and for various reasons, this might use formats more aligned to

rapid comparison such as the number of seconds since 01-01-1970 https://en.wikipedia.org/wiki/Unix_time.

Regards

Pete Rivett, Principal, Federated Knowledge LLC
pete.rivett@federatedknowledge.com