REGULATORY GENOME DEVELOPMENT LTD (TRADING AS REG-GENOME)
Kett House, Station Rd, Cambridge, United Kingdom, CB1 2JH
reg-genome.com

Cambridge, 21 February 2023

Office of the Corporate Secretary, FINRA

1735 K Street, NW

Washington, DC 20006-1506

Attn of Ms Jennifer Piorko Mitchell

Dear Madam

Special Notice 211022: RegGenome Comment on FINRA's Machine-readable Rulebook Initiative

1. Background

RegGenome is pleased to respond to the questions raised by FINRA in the above publication.

We are a provider of structured regulatory content, spun out of the University of Cambridge in 2021 with a mission to transform how the world produces and consumes regulatory information. As a co-founder of the Regulatory Genome Project, we work with the University as well as regulators to help build and promote the adoption of an open, jurisdiction-agnostic framework for representing regulatory obligations. This framework is called the Cambridge Regulatory Genome (CRG) and we believe it can revolutionise the development of RegTech applications as well as empirical research.

RegGenome uses natural language processing (NLP) and machine learning to identify and tag obligations from the CRG ontology in regulatory text, representing regulatory information as structured electronic content that is portable and interoperable across jurisdictions, functions and applications. By mid-2023 our structured content will cover the majority of regulated financial services across the majority of jurisdictions in the world. We provide the content that powers the University's digital tools for Regulators and work with regulatory as well as legal experts to build reliable information structures – the CRG taxonomies.

Through our collaboration with the University of Cambridge, we are actively working with standard-setting bodies to explore how the CRG taxonomies can be used in capacity building and self-assessment; and moreover working with stakeholders across the regulatory landscape to validate and improve the CRG taxonomies as a public good.

2. Making the most of FIRST

We have been honoured to observe the development of FIRST since its early days and are pleased to see it released to the market as an excellent tool. In terms of how

information is organized in FIRST, we find that the current structure is intuitive and very valuable, with combined search terms, in particular, driving very precise results.

Given the potential value of FIRST itself and the resource that has gone into it, we have framed this response as a discussion of how it can achieve broad adoption and be made future-proof to the extent possible. Ultimately, we believe that the promise of such tools can best be fulfilled by

- Making both content and information structures easily available for both compliance use and commercial re-use in appropriate, widely recognised formats.
- Focusing the work of regulators and self-regulatory bodies on what they are uniquely well-placed to do compared to private firms
- Creating open information structures that can be built upon by other parties within the regtech value chain, under terms that ensure responsible use.

3. Delivering content effectively

The highest-value use cases for ontologies and data such as those underlying FIRST are unlikely to involve searches via a Graphic User Interface – rather they will involve queries from one compliance application to another, or to a Golden Source of structured regulatory data.

Delivery via an API is definitely more valuable to RegGenome as a content vendor than the alternatives; and more so if FINRA intends to expand FIRST. More generally, web scraping, file drops and other similar approaches are not ideal for either producers or consumers of regulatory information. API access gives the regulator greater control, and provides vendors and firms with much higher quality content, including the ability to add crucial contextual information such as flags to indicate that content as superseded or has been corrected elsewhere. These options in turn reduce errors and adoption costs and also ensure that parties that have obtained content through unsanctioned web scraping will not be able to disguise the resulting content as authentic. The ability to bring as much as possible of the tagging that went into FIRST into the API will greatly enhance its value, even if only a few rules are tagged in this way.

It is important that as FINRA focuses on perfecting API delivery of the FIRST data, some degree of attention is also paid to the terms under which such content is available. Pure-play content providers such as RegGenome rely on distribution channels in order to get a product to market, meaning that terms restricting distribution to vendors with their own platforms or with direct contracts with end-users will unnecessarily constrain the market without much benefit, in our view, to FINRA's objectives.

4. Creating open information structures

Regulated firms use technology extensively in compliance; yet each vendor they onboard brings their own ontologies which must then be reconciled with the hierarchies of policies, controls, datapoints, functions and risks that a firm already adopts internally. This is crucial for information to flow within compliance systems. This

of course is costly to all parties and, at the macro-level, wasteful. Adoption of any new structure, such as a taxonomy, therefore proceeds faster if vendors are allowed to pre-map their products to regulators' ontologies and expand on the latter in order to represent the firm's internal universe of policies, controls, data and functions.

FINRA may need to allow for this specifically in devising a content license for FIRST-enabled data, e.g. by obliging re-users to state the version of the taxonomy that they are using and any changes they've made to it. It is particularly important that users do not, for example, insert into their literature claims of 'mapping to FINRA's ontology' without adhering to at least some standards of responsible adaptation.

We believe there is a role for jurisdiction-agnostic ontologies such as the Cambridge Regulatory Genome to further support FINRA in driving adoption. Most major firms are international and will likely have operations beyond N. America, to which they will likely want to apply internally-consistent compliance policies. While FINRA cannot be expected to maintain ontologies for all potential jurisdictions of interest, mapping the FIRST ontology to the CRG and vice versa would allow third parties to build enhanced functionality such as 'nearest equivalent rule' lookups across jurisdictions. Such functionality will be particularly important to firms seeking to maintain consistent policies across their global operation.

Finally, if FINRA does seek to drive adoption by incremental adaptation, then it is important that documentation is provided that would ensure users understand (and developers do not abuse) the boundaries of different concepts. E.g. ACC, BU, FT and RP classifications should be accompanied by advice on how to annotate these, or at the very least guidance on how to distinguish categories from one another – e.g. Capital Markets from Trading and Execution.

Even if no third parties create adaptations, the FIRST taxonomy will evolve and it will be useful that FINRA maintain it in a public, version-controlled way that allows as much backward compatibility as possible so that changes do not result to breaks in workflows for users that have embedded the structure into their workflows.

5. Building on the regulator's strengths, and understanding those of others

Certain FIRST features cannot be easily replicated by private firms and provide potential areas for FINRA to focus further development on. We would single out entity recognition (in this context, ascertaining what type of firm the content creates obligations for and tracking that same type of firm across segments of text or even documents). Entity recognition is a very difficult challenge for NLP-based classification models, because language specific to an entity type is typically not present in every piece of text that affects those entities. Moreover, a regulator has a very clear advantage over any third party in providing entity metadata as they are an authoritative source of perimeter guidance. The value to firms of knowing what applies to them and what does not is, needless to say, particularly high.

In terms of future expansion of FIRST to further FINRA rules, we feel that FINRA's decision to focus on the relatively few high-demand rules was correct at this stage and

outward expansion should be slow and tightly controlled. FINRA should instead encourage the development of classification models for the documents or areas of its ontology that it cannot manually label itself and release its own annotations and guidance as inputs into this process (ie seed training data). While third party classifications might never be endorsed by FINRA, a standard for when a classification model may (not) be said to be aligned to FINRA's ontology should be put in place.

The ability to navigate the rulebook by similar terms (as raised in the Call for Evidence) is potentially interesting. However we would warn against superficial or coincidental connections. The ability to navigate by regulations, rules and laws (RL) is likely a more promising start and term-driven navigation might best be implemented as subordinate to and supporting of such RL-based search; this would limit the number of false-positive results significantly if FINRA were to rely on machine-driven predictions to further grow FIRST.

6. Other features for prioritization – RegGenome's perspective

In preparing our response, RegGenome has explored how well FINRA's topic classifications map to the Genome and our upcoming taxonomies for Capital Markets regulation. In our view, not all of FINRA's Detailed Topics are more valuable than summary topics in this regard. This is because summary topics appear to map better to firm-side artifacts with clear use cases, such as policies, processes, controls, or obligations. Aiming for a similar level of user-case relevance would help improve the current Detailed Topics taxonomy. Moreover, it is not as clear to us how terms were selected for RegGenome Terminology and Defined Terms nodes in the detailed taxonomy. In our view there is great value in concentrating on defined terms that define the application of rules or the perimeter of the regulatory or legal framework for an activity – in time such terms can be paired with questionnaires or other diagnostics that allow firms to self-classify, thus customising their view of the rulebook based on what rules apply to them.

Clarity on these matters, and ideally published guidance for users, will allow FINRA to more confidently expand coverage of detailed topics. Typically obligation- or -control level granularity is harder to accommodate and more reliant on guidance, because of the level of interpretation involved and the diversity of the firms subject to the rules. In our experience, there may even be value in a regulator setting out a maximum level of granularity that they are will to engage in, so that industry and vendors can assess what work they need to do in order to integrate with the regulator's ontology. This handover from authoritative ontologies to industry consensus-based ones is important because it allows regulators to construct useful ontologies without being forced to provide industry guidance in the process.

It is worth noting that any vendor such us ourselves will benefit the most from data elements that are orthogonal to those in their own ontologies – e.g. the CRG features obligations-based taxonomies, which will additionally be mapped to business functions. It is therefore elements most closely linked to entity recognition and information

extraction (security type, firm type, customer type, numerical reference) that would be most valuable from our perspective. A risk taxonomy is also likely to provide a strong complement, and moreover a good way of mapping FIRST further against firms' internal ontologies. These however are expectations based on our own experience. A stocktake of how vendors organize information will be useful to FINRA in planning next steps in the expansion of FIRST.

Despite our reservation above, expanding coverage to more rules will clearly be more beneficial to vendors such as ourselves than deepening an ontology which is already quite comprehensive. The latter approach would, in our view, risk slowing down the broader initiative behind FIRST by increasing the degree of complexity.

Conclusion

We are excited about the potential of FIRST – we see the promise of significant savings in quality assurance of documents and synergies in combining FINRA's classification with others in the industry, including the Cambridge Regulatory Genome. We also see great value in delivery that does not depend on scraping or file drops. We do however hope that FINRA will consider how it can allow for redistribution models such as RegGenome's, so that pure-content players are not disadvantaged vis vendors who run their own platforms. We will be happy to discuss any of the above matters with FINRA and the FIRST project team as well as explore the potential for integration between our respective ontologies.

Yours sincerely,

Emmanuel Schizas

Head of Taxonomy Development

Regulatory Genome Development Ltd.