Financial Entity Identification and Information Integration (FEIII) Challenge: The Report of the Organizing Committee

Mark Flood
Office of Financial Research
mark.flood@ofr.treasury.gov

John Grant
University of Maryland
grant@cs.umd.edu

Haiping Luo
Office of Financial Research
Haiping.Luo@ofr.treasury.gov

Louisa Raschid
University of Maryland
louisa@umiacs.umd.edu

Ian Soboroff
NIST
ian.soboroff@nist.gov

Kyungjin Yoo
University of Maryland
athleta@umd.edu

ABSTRACT

This report presents the motivation and challenge tasks for Year One of the Financial Entity Identification and Information Integration (FEIII) Challenge. It summarizes the process and outcomes as well as lessons learned and future plans.

Keywords

FEIII, record linkage, named entity resolution, LEI.

1. MOTIVATION AND INTRODUCTION

The Year One goal of the Financial Entity Identification and Information Integration (FEIII) Challenge was to develop technologies to automatically align diverse financial entity identification schemes and identifiers. This will be an important first step in developing a reference financial entity identifier knowledge base linking heterogeneous collections of financial entity identifiers. Such a (Linked Data) knowledge base will help to overcome a fundamental roadblock to successful information integration, i.e., the resolution of mentions or references to the same financial entity across the financial landscape of financial contract prospectus, regulatory filings, transactional records, news articles, social media, etc.

Named Entity Resolution / Entity Resolution (NER / ER) is a well studied problem. Challenges and solution approaches for NER / ER, in general, and the particular characteristics relevant to the financial domain, have been presented in [2, 3, 4, 6, 8, 10, 11].

The tools and knowledge bases developed around the FEIII Challenge should enhance the toolkit of researchers, industry participants, and regulators, who must merge and align data from a broad array of sources, including financial firms' internal systems, regulatory collections and public websites. This is an ongoing, repeated task, and one that does not scale easily. Tools to improve the efficiency and accuracy of

these data alignments will be an important building block of a more resilient financial system.

The FEIII Challenge is envisioned as a series of challenges that are being coordinated by the University of Maryland (UMD), the National Institute of Standards and Technology (NIST) and the Office of Financial Research (OFR). Co-organizers Louisa Raschid, Ian Soboroff and Mark Flood have been collaborating with a team of domain experts from the financial industry and federal regulatory agencies, and with computer and information scientists, for over a year, to identify FEIII Year One challenge problems, data, ground truth, etc.

We provided entity data from three sources, the Federal Financial Institutions Examination Council (FFIEC), the Securities and Exchange Commission (SEC), and the Global Legal Entity Identifier Foundation (GLEIF). Each source included its identifier code as well as identifying features such as names and addresses. The data varied from source to source in content, completeness, and quality.

The Year One challenge task was to align entities across the sources, i.e., to match FFIEC identifiers to LEIs and to match FFIEC identifiers to SEC identifiers. Participants returned proposed matches as identifier pairs. This process was facilitated with a simple website that allowed participants to register for the challenge, to receive a pointer to the data, and to submit their results. We computed precision and recall scores against a ground truth dataset of matches¹. Details of how the ground truth dataset was created by experts is provided in the paper.

A total of eight organizations completed the challenge. These organizations included both academic and industry sectors and came from both the US and Europe. All participants completed the first two required tasks and a few participants attempted the next two optional tasks.

The report is organized as follows: Section 2 presents details of the challenge task, the dataset and the ground truth dataset. Section 3 presents the (self reported) approaches taken by the participants and summarizes the scores. Section 4 discusses lessons learned and future plans.

2. YEAR ONE CHALLENGE

¹Ground truth is a term that is used to refer to information provided by direct observation as opposed to information provided by inference. In our context, the ground truth are the matches that are validated by experts.

2.1 Task Overview

We distributed the following data files to challenge participants:

- **FFIEC.csv**: This input (file) from the Federal Financial Institutions Examination Council (FFIEC) provides information about banks and other financial institutions that are regulated by agencies affiliated with the Council. The primary key identifier is labeled **Research Statistics Supervision Discount ID (RSSD ID)**.
- **LEI.csv**: This input from the Global Legal Entity Identifier Foundation (GLEIF) contains the **Legal Entity Identifier (LEI)** for a wide range of institutions.
- **SEC.csv**: This input from the Securities and Exchange Commission contains entity information for entities registered with the SEC. The primary key identifier is the **Central Index Key (CIK)**.

These files were distributed in both comma separated values format and as Microsoft Excel `.xls` spreadsheets. A fourth file, **FHLB.csv** from the Federal Housing Finance Agency, listing member institutions of the Federal Home Loan Bank System, was distributed but was not used in a scored challenge task. The data distribution also included a data dictionary, a small set of sample matches, and a background slide presentation.

The files all differed with respect to the attributes (columns) as well as the quality and consistency of data. We did not try to *harmonize* values across these files. This was a conscious decision on the part of the organizers to mimic how the data typically arrives from an institution in the regulatory world.

The data dictionary was descriptive but did not account for data errors. Inconsistencies in the data included whether institution names were abbreviated, word ordering in names, address formats, and missing data. We note that all inconsistencies were present in the data received from the data owners; no inconsistencies were introduced.

- **Task 1**: Align records between the FFIEC and LEI files using only the name and address data provided.
- **Task 2**: Align records between the FFIEC and SEC files using only the name and address data provided.

By *align* we mean that challenge participants were to produce identifier pairs (a, b) indicating that identifier a from the first file refers to the same institution or entity as identifier b from the second file. Thus, for Task 1, participants produced a two-column file, where the first field in each line was an FFIEC identifier, and the second the LEI, for an organization that the participant was claiming to be the same entity as represented by the FFIEC identifier.

Note: There were four planned tasks, but budget and timing constraints only allowed us to create ground truth datasets for Tasks 1 and 2. Task 3 was to align identifiers across the three files above, and the fourth was to align LEI and SEC identifiers. We plan to discuss Tasks 3 and 4 and release some results at the workshop.

2.2 Constructing the Ground Truth

We constructed the ground truth by first generating candidate matches, and then having a team of up to three experts

adjudicate the candidate matches. We used the Duke probabilistic matching tool [7, 9] to generate candidate matches. Since we planned on manual adjudication by experts, we did not focus on tuning the Duke tool to produce high-precision candidate matches; instead we used Duke as a recall-oriented filter that sampled the data for manual annotation.

We presented the candidate matches from the Duke tool to experts at the Office of Financial Research. They labeled the matches as correct (True Positive), incorrect (True Negative), or ambiguous (Ambiguous). While in many cases matches were trivially correct or incorrect, there were a number of cases in which the experts needed to rely on domain knowledge as well as external resources. These latter cases illustrate the importance of the FEIII Challenge. Where the experts disagreed, we resolved the judgment through discussion among the experts, as well as through consulting additional resources. These included the National Information Center (NIC) database from the Federal Reserve Board [5] and the Alacra Authority File and Resolve database [1]. Matches that were not resolved were labeled to be Ambiguous.

We obtained a second pool of candidate matches by *chaining* over the RSSD ID identifiers from the True Positive manual judgments, for the two tasks. Additionally, we pooled candidate matches from the challenge submissions that had not been manually adjudicated. We asked the expert judges to label a sample of both these pools.

This process produced a ground truth dataset that was not complete, i.e., we have not manually adjudicated every candidate match. However, we claim that the ground truth is sufficiently complete so that we can measure participant performance on the tasks.

Table 1 provides statistics of the ground truth for both tasks. The first row is the count of FFIEC identifiers from the source file. The middle rows are counts of matching pairs and these counts may contain duplicate matches for some FFIEC identifiers. The last row reports on the count of FFIEC identifiers that were not included in any of the matching pairs (True Positive or Ambiguous or True Negative), and were not adjudicated by the experts for Tasks 1 (or 2). This last count represents the incomplete nature of our ground truth dataset. The limitations of an incomplete ground truth dataset will be discussed in the next section.

For scoring the challenge submissions, we dropped the entries labeled ambiguous and computed precision and recall. Precision is defined as the fraction of matches submitted by a participant that were true positive matches according to the ground truth. Recall is defined as the fraction of all known ground truth true positive matches submitted by the challenge participant. We computed the F-score, a balanced ($\beta = 1$) harmonic mean of recall and precision. We note that for the majority of participants, all of their entries were manually adjudicated by one or more experts.

3. SOLUTION APPROACHES

We asked challenge participants to briefly describe their approach and to indicate any resources used outside the data provided within the challenge. The following are the **verbatim** text of their descriptions in no specific order. We note that participants have been invited to provide their own reports to be discussed at the SIGMOD Workshop on Data Science for Macro-Modeling with Financial and Economic Datasets (DSMM 2016).

	FFIEC	to-LEI	to-CIK
FFIEC count	6652		
True Positive pairs		496	230
Ambiguous pairs		38	95
True Negative pairs		932	560
FFIEC unadjudicated count		5419	6011

Table 1: Summary statistics for the ground truth for Tasks 1 and 2. The first and last rows are counts of FFIEC identifiers while the middle rows are counts of matching pairs.

- FactSet Research Systems, Inc.
External resources: Only those internal to FactSet, such as the unique identifier matching algorithms and entity concordance methodologies.
Approach: Compiled the files into separate SQL tables and then ran comparisons between those tables matching names and locations.
- Hasso Plattner Institut - Information Systems
External resources: the rule were extracted from the datasets and some external resources, such as: Google Finance, Bloomberg, Reuters, Bank SWIFT (BIC) codes.
Approach: The final result is a combined solution of hybrid similarity measures for attributes like the company name, extracting parts of the addresses (like floor, suite, po box etc.) and applying some rules to prune record pairs, which we were able to derive from the ground truth plus from the datasets.
- IBM Research
External resources: none.
Approach: used method based on High-level Integration Language (HIL), which is part of IBM’s suite of entity analytics.
- ISMB & EURECOM
External resources: Lexicon.
Approach: based on Duke, with an ad-hoc developed cleaner and a final pruning step based on lexicon (threshold 1).
- Pennsylvania State University
External resources: none.
Approach: 2-step matching is done for record linkage 1. Exact matching after cleaning records(address, name) 2. Random Forest (RF) based matching: used features with various different string comparison distance of name, street address, city, state, zip code
- Tahoe Blue Ltd.
External resources: none.
Approach: the results produced by the network analysis utilized scores that represent the likelihood of matches among entities by iterating over pairs of the datasets. The results are interesting in that the thresholds on match scores for considering high-confidence matches can be adjusted, and - as you can see from the graph visualization - there are a few cases where some alternative candidates for identity integration met the threshold criteria.
- Thomson Reuters
External resources:

FFIEC to LEI	Precision	Recall	F-score
Minimum	17.62	77.02	29.13
Maximum	99.24	96.37	97.44
Mean	80.49	86.90	79.78

FFIEC to CIK	Precision	Recall	F-score
Minimum	10.13	56.52	17.88
Maximum	92.82	85.65	88.38
Mean	63.86	71.80	62.36

Table 2: Minimum, Maximum and Mean values for Precision, Recall and F-score for Tasks 1 and 2.

Approach: the approach has relied upon fairly robust analysis of the informational elements in reference data without seeking to semantically parse the nature of the entity from its name.

- University of Texas at Austin
External resources:
Approach: block purging+thresholded Jaccard similarity on select fields, after tokenizing

Table 2 provides statistics of the scores for both tasks. Participants were provided with their individual scores and summary statistics. A ranking of participants by the F-score will be provided at DSMM 2016.

4. LESSONS LEARNED, IMPACT AND FUTURE PLANS

The Year One Challenge is a first step to develop tools to semi-automatically align diverse financial entity identification schemes and identifiers. In particular, the challenge was limited with respect to the real world problem of identifier alignment. Consider that there are thousands of identifier schemes, each developed by different organizations. Such schemes were defined due to the absence of, or slow adoption of, global identifiers such as the LEI, or due to a specific need in a market sector. Aligning identifiers across a limited number of schemes, with the traditional approach of pair-wise matching and computing precision / recall scores, is straightforward. Unfortunately, it will not scale well with hundreds or thousands of schemes. While we do not address scalability, it is an important motivation driving the development of semi-automated approaches to alignment.

This year we created a small ground truth dataset. This dataset was sufficient to measure the performance of the participants in the challenge. From the last row of Table 1, we observe that approximately 20 % of the FFIEC identifiers were manually adjudicated for Task 1, while approximately 10 % of the FFIEC identifiers were manually adjudicated for Task 2. This leaves a significant completeness gap where 80 % to 90 % of the FFIEC identifiers were not adjudicated over these two tasks.

One of the lessons learned is that the incomplete nature of the ground truth will remain a concern for many reasons. For example, future uses of the ground truth include the ability to use it as training data, to tune sophisticated new solutions. It is possible that future systems could then produce a significant number of unadjudicated match candidates. Without more complete ground truth, we would not be able to determine the performance of these systems

and distinguish good solutions from poorly performing ones. Users of the ground truth dataset should thus be cautious in situations where there are many adjudicated candidates matches, and / or when making comparisons to systems that participated in the challenge.

Another lesson relates to the ambiguous matches in the ground truth. These are candidate matches that our expert team could not resolve satisfactorily. On the one hand, ambiguity validates the importance of the FEIII challenge, i.e., there exist pairs of entity identifiers that cannot be aligned reliably by experts, much less by algorithms. On the other hand, ambiguous matches are a limitation in the ground truth dataset. We cannot reward systems or punish them when the matches are ambiguous. If the count of ambiguous matches increases, then this will also prevent us from correctly identifying the best performing solutions.

There were several limitations to our Year One Challenge that will lead to interesting Year Two Challenge tasks. One limitation was that the identifier lists may be incomplete. They also evolve over time so we were not always attempting a synchronous match. Temporal evolution also means that these matches must be kept updated.

A more important limitation is that we were matching sets of identifiers obtained from primary resources that were able to provide values for names and addresses. The more interesting challenge is NER / ER of *financial entity name mentions* in news sources, Web pages, financial contract prospectus, transactional records, etc. Related research on addressing NER / ER of legal entity name mentions from residential mortgage backed securities prospectus filed with the SEC has been reported in [2, 11] and will inform our Year Two Challenge.

Another interesting limitation was that we addressed primarily a one-to-one matching problem, i.e., we expected each RSSD ID identifier to match a single LEI, or a single CIK. In the real world, a financial entity name mention may potentially match a large number of LEI (CIK) values. The challenge there would be to first determine the (unique) correct match, and then to determine the *semantic relationships* among other candidate LEI (CIK) matches, e.g., a parent, subsidiary, national association, etc.

Despite these limitations, we believe that the FEIII Challenge series, even in its first iteration, could have a significant impact. The ground truth dataset that was created is an open resource that can be used to benchmark the performance of existing commercial identifier matching systems. It will provide training data to improve existing services. More importantly, the FEIII Challenge was the first successful attempt to define a protocol for NER / ER tasks in the financial domain.

An aspirational goal is that every document would eventually be associated with all relevant LEIs. In that context, the FEIII Challenge could lead to the development of systems that would annotate any document with a set of LEIs, determine all possible relationships among these LEIs, and provide a confidence score for the predicted annotations and relationships. Another future service could align identifiers from different points or intervals of time, providing a historical context for current identifiers.

Finally, the organizers hope that the FEIII Challenge will help foster a community of researchers from computer science, finance, economics, and mathematics, representing academia, industry and the regulatory community, all with a shared

interest in data science for finance and financial big data challenges.

5. ACKNOWLEDGMENTS

We thank the following individuals for feedback on the challenge tasks: Robert Dunn, Greg Feldberg, Craig Knoblock, Bill Nichols, Ishwarya Sethuraman, Stathis Tompaidis, Maria Esther Vidal and Julie Vorman. The following individuals provided administrative and logistical support: Evelyn Brown, Ed Garnett, Walter Hamscher, Steve Olden and Bill Ruberry.

This research was partially supported by awards NIST 70NANB15H194 and NSF CNS1305368, and the Smith School of Business.

6. ADDITIONAL AUTHORS

Elena Zotkina (University of Maryland, email: ezotkina@umiacs.umd.edu).

7. REFERENCES

- [1] Alacra resolve entity lookup. <https://www.alacra.com/reference-data-solutions/lookup/>.
- [2] D. Burdick, S. De, L. Raschid, M. Shao, Z. Xu, and E. Zotkina. resMBS: Constructing a financial supply chain graph from financial prospecti. In *Technical Report, University of Maryland*, 2016.
- [3] D. Burdick, M. A. Hernández, H. Ho, G. Koutrika, R. Krishnamurthy, L. Popa, I. Stanoi, S. Vaithyanathan, and S. R. Das. Extracting, linking and integrating data from public sources: A financial case study. *IEEE Data Eng. Bull.*, 34(3):60–67, 2011.
- [4] L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss, and S. Vaithyanathan. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1002–1012. Association for Computational Linguistics, 2010.
- [5] National information center. <https://www.ffiec.gov/nicpubweb/Content/HELP/HelpAboutNIC.htm>.
- [6] M. Flood, H. Jagadish, and L. Raschid. Big data challenges and opportunities in financial stability monitoring. *Financial Stability Review, Banque de France*, 20:129–142, 2016.
- [7] L. Garshol. Bayesian identity resolution. <http://www.garshol.priv.no/blog/217.html>.
- [8] L. Getoor and A. Machanavajjhala. Entity resolution: theory, practice & open challenges. *Proceedings of the VLDB Endowment*, 5(12):2018–2019, 2012.
- [9] Duke github repository. <https://github.com/larsga/Duke>.
- [10] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Linguisticae Investigationes*, 30(1):3–26, 2007.
- [11] Z. Xu, D. Burdick, and L. Raschid. Exploiting lists of names for named entity identification of financial institutions from unstructured documents. *arXiv preprint arXiv:1602.04427*, 2016.