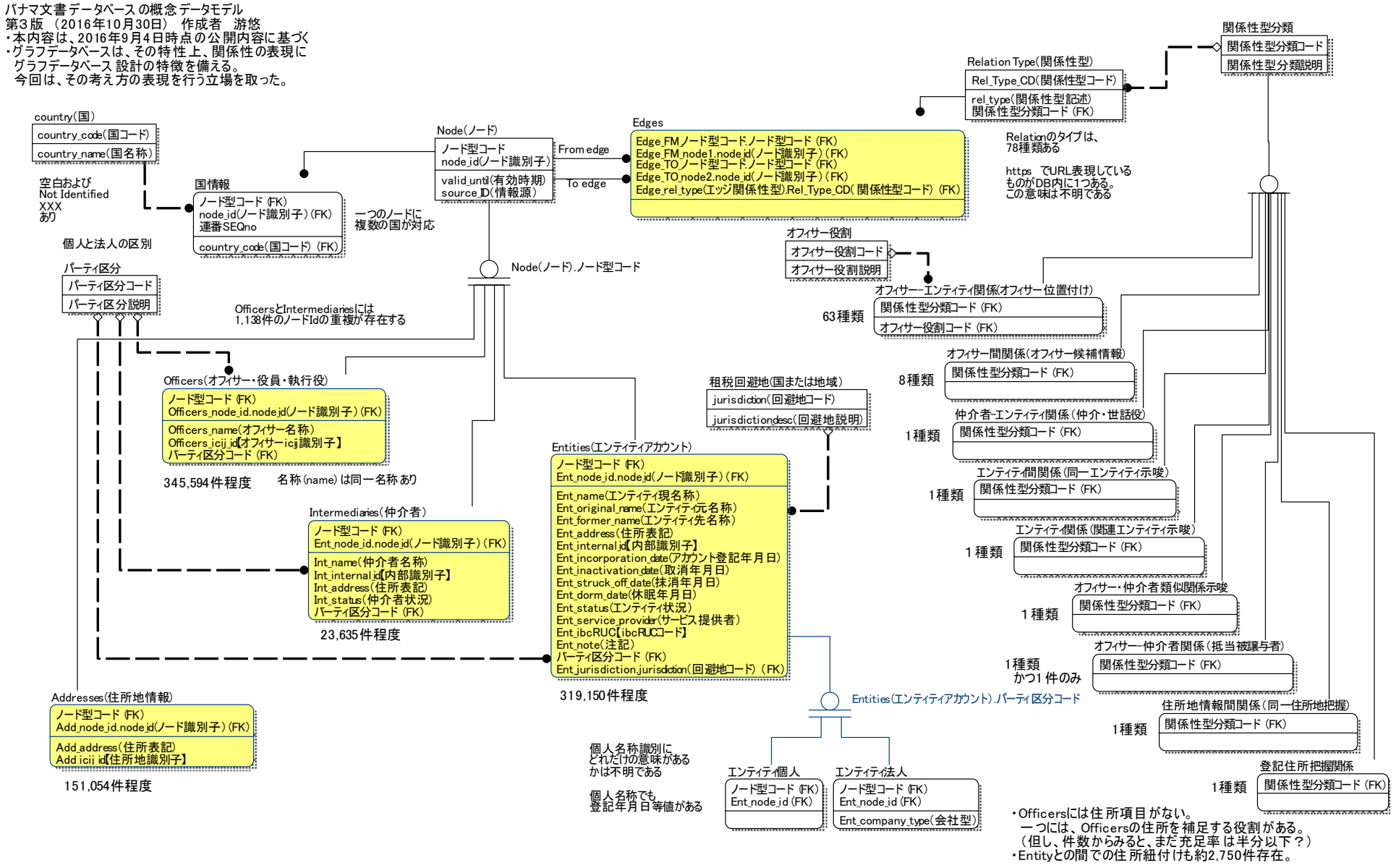
【参考1】Panama文書DBデータのER構造図（2016年9月4日時点内容を元にした）



●以下の内容は、JEMUGのモデル解説へ筆者が記載した内容を転用したものです（一部修正）
元のデータモデリングコンテストの出題内容は、以下のJEMUG　Webページを参照下さい。

https://jemugguestroom.wordpress.com/　　2016年9月21日記事（2016データモデリングコンテスト募集要項）

モデル作成のポイントについての解説 Written by 游悠 (2016.12.13)
今回授賞したデータモデル作成時の考慮ポイントについて簡単に説明します。利用したモデル作成ツールには、ERWinのコミニュティエデションを使用しているため、25エンティティ以内でコンテスト応募ERモデルを作成するという条件があることも考慮条件でした。

(1)基本エンティティの取扱いについて
まず、Panama文書DB情報（今回csvファイルとして提供）の基本エンティティとして、Officers、Intermediaries、Entities、Addressesの４つを取り出すことができます（以下、それぞれ「O」、「I」、「E」、「A」と略す）。これらはデータ特性からノードIdがキーとなっていることが分かり、元々利用されているグラフDB（Neo4J）との関係からみて、ノードというエンティティのサブタイプとして扱えることになりました。
そこで確認が必要なのは、ノードId単独でO、I、E、Aの各エンティティの全データがユニークに識別できるかということです（つまり、ノードエンティテイの主キー（PK）として、ノードId単独項目が使えるかということ）。データ調査の結果、OとIのノードIdには1,100件を越える値の重複があることが分りました。この件数は単なる入力誤りなどによるものでなく、DB作成者の意図によるものと考えられます（例えば、OとIの中の出現者（人または企業）は、時々の果たす役割で使い分けられている）。
そこでノードエンティティの主キーには、ノード型コードの追加が必要であると結論付けをしました。そして基本４エンティティの共通属性としてvalid_until、source_ID、country_code（国コード）の３項目を取り出せます。この中で国コードは、データを調べると１件当たりに複数の国を記述していることが分るため、CountryとNodeとの関係として取り出しました。これは単なる多対多として表現できますが、ここでは記入順番が判断の優先度と係わる可能性も考慮し、連番で識別できる形で表現しています。valid_untilとsource_IDはノードエンティティへの属性項目として記述し、モデル全体エンティティ数制限の関係からモデルエンティティとしては表現せず、コード値についてコメント内に補足をするに止めました。

(2)E、I、Oエンティティについて
これらのエンティティ内データには、企業名と個人名が混在するため、これらを区別する意図としてパーティ区分エンティティを切り出しました。これは特にEのcompany_type項目の位置付けに注目させる表現として役立つと考えています。Eについては、jurisdiction（租税回避地）がモデル表現理解の上で意味を持つという考え方から、租税回避地エンティティとして見える形で表しています。
また、Eエンティティの日本語名称は、「エンティティアカウント」と名称表示しました。その理由として、エンティティの属する各レコード（ロウ）の開設日などの項目を設けた視点を取り入れ、取引口座という解釈を与えるためであることを付記します。

(3)Aエンティティについて
住所項目はIとEには項目として存在し、Oには存在しません。そしてこのAが別途のエンティティとして元のDBで扱われています。そしてこれに係わる関係性がリレーションとして使われています。このため、Aエンティティの持つ住所情報は、DB作成者のデータ利用の意図を表すと考え、IとEでのaddress項目はレコードに関する「住所地表記」として項目命名し、またAエンティティは「住所地情報」として名称付けを行い、名称によりそれぞれの位置付けを区別するモデル表現にしました。

(4)Edgesエンティティと関係性（リレーション）の表現について
パナマ文書情報がグラフDBを用いて開示されていることの大きな目的の中で、この関係性表現を理解することが大切であると筆者は判断しました。このため、Edgesのcsvファイルで提供される関係性情報を今回のデータモデルでもきちんと表現するのが重要と考えました。それがNode間の関係としてEdgesエンティティを記述した理由です。Edgesエンティティの主キーには、リレーションタイプ（関係性型）を含めています。この関係性型は、元のパナマ文書DBではグラフ上のエッジ（有向辺）として示される関係性（78種類（※1））を、その関係性の果たす役割（繋いでいるノードの型や意味合いなど）によって9種類に筆者が分類したものです。この関係性型の中で最も属する種類が多かったのが「オフィサー-エンティティ関係(オフィサー位置付け)」で、63種類ありました（例えば、取締役、株主など）。そこでこの関係種類をオフィサー役割として取り出し、エンティティ表現したものです。上記の9分類を関係性型分類エンティティのサブタイプ化しました。

以上のような考え方で、全体25エンティティというツール制限一杯で表したものが、今回の授賞したER図です。

注 ※1 元のDBではユニーク値は80種類となりますが、例えば「Shareholder of」と「shareholder of」という具合に同一として扱えると判断できるものがあり、それらを一緒にして、78種類あるとしました。

●補足

筆者は、このパナマ文書DB（グラフDB）のデータが表す全体像が、ER図表記だけでは捉えにくいという考え方をして、別途ネットワークグラフ表現を用いた、全体概念の様子を示す概念図も作成しました。それを以下、参考として付記します。

【参考2】 Panama文書DBデータのカバー構造イメージ図

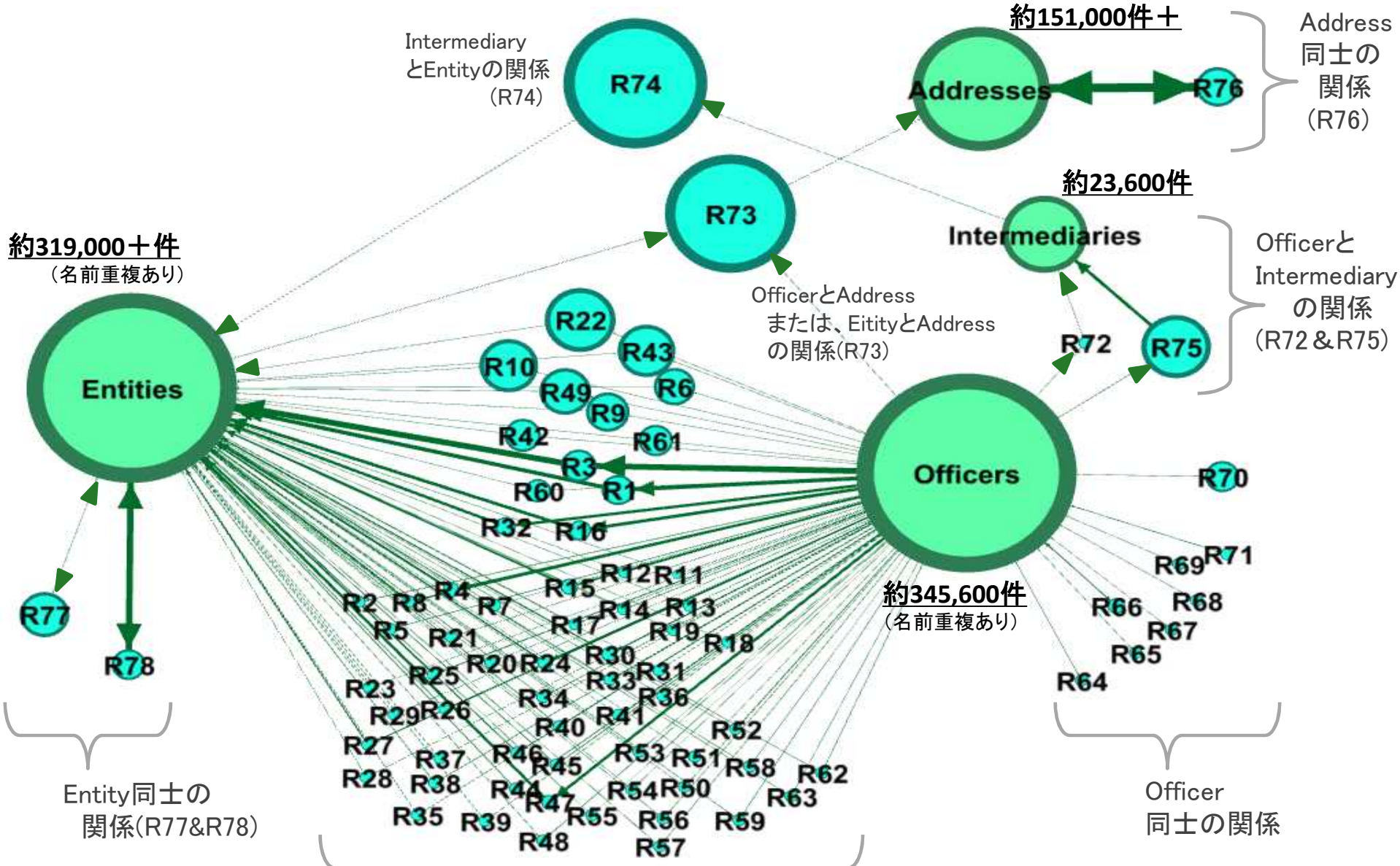
・緑色のサークルが主要4エンティティを示す。
・青色のサークル（円）が関係を表すノードであり、その円の大きさでおおよそのリレーション数を示した。これにより当該リレーションを備えるインスタンス数がどのように分かれているかを見ることができる。「Rnn」はリレーションの種類を示す識別子。この内容は別途の表で示しているが、ここでは省略する。R73、およびR74が大きな関係者数を表していることが分かる。

【参考3】 Panama文書DBデータのカバー内容イメージ

・参考2の図を、更に関係者数の重なり数の大きさを踏まえてイメージ化した図である。これにより、関係者の全体の大きさの相対的イメージを理解することができる。

【参考2】Panama文書DBデータのカバー構造イメージ図（2016年9月4日時点内容を元にした）

（ノード円の大きさが概要の相対度数の大きさを示す）



備考：

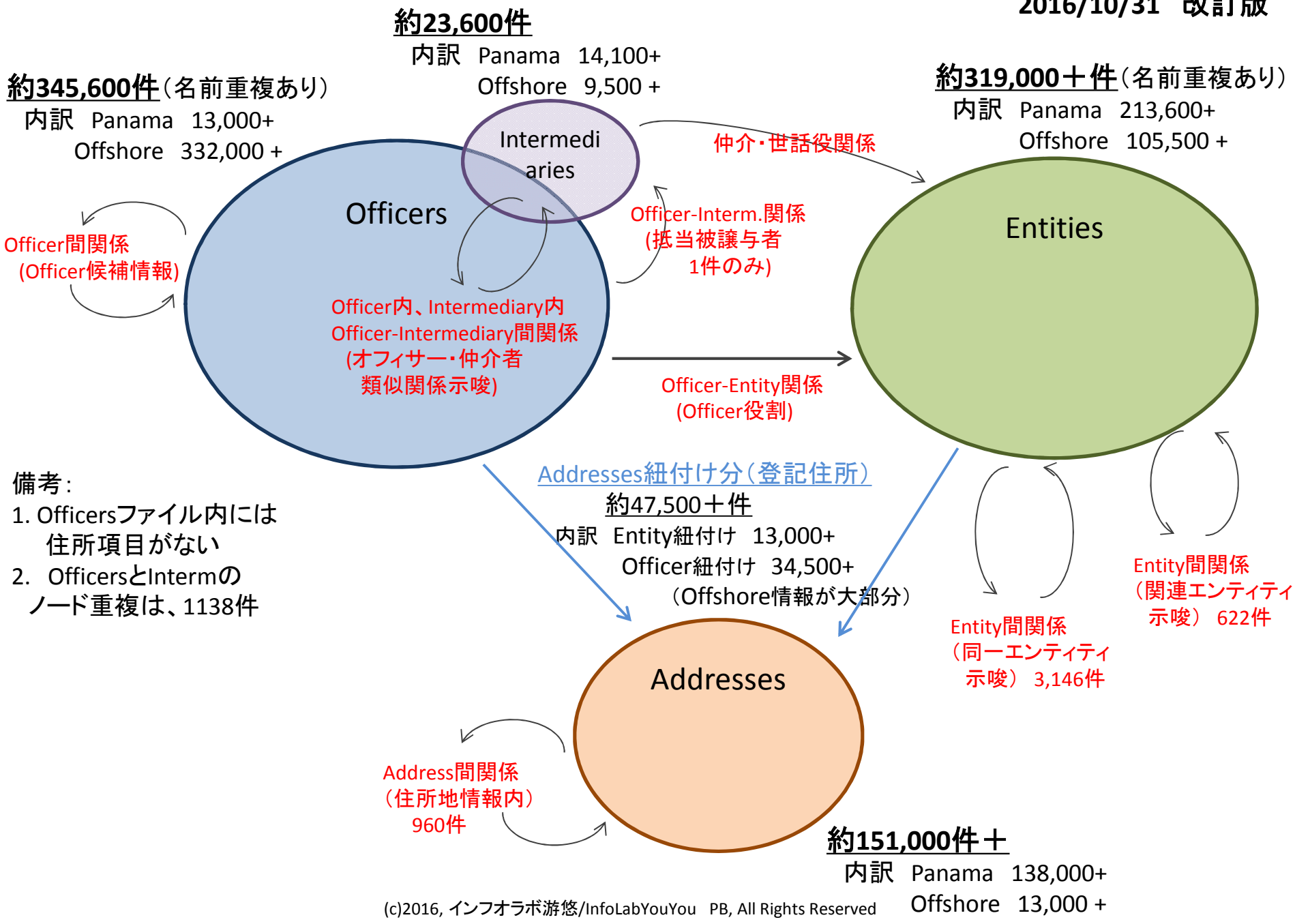
1. Officersファイル内には住所項目がない
2. 実際にはOfficersとIntermediariesの ノード重複が、1,138件ある

凡例：・「Rxx」名称は、リレーション(内容は別表参照)
・Officers,Intermediaries,Entities,Addressesは基本のエンティティ

2016.11.18

【参考3】Panama文書DBデータのカバー内容イメージ（2016年9月4日時点内容を元にした）

2016/10/31　改訂版



備考：
1. Officersファイル内には
住所項目がない
2. OfficersとIntermの
ノード重複は、1138件