

# ドメイン固有オントロジーを用いた 時系列データからの要約文の自動生成

近藤 颯<sup>1</sup> 沼尾 雅之<sup>1</sup>

**概要**：近年，IoT の普及によりセンサなどから得られる時系列データは様々な場面で利用されているが，データをそのまま表やグラフで表示しても専門知識のない人が解釈することは非常に困難である．そのため，時系列データの特徴を要約した文章を自動生成する研究が行われている．しかし，多くの情報を含んだ時系列データからの文章生成にはそのドメイン固有の知識を用いることが必要である．本研究では，ドメイン固有のオントロジーを用いることでドメイン固有の知識を反映させた時系列データからの要約文生成システムを提案する．任意の時系列データから，そのドメイン固有のオントロジーを構築し，時系列構造オントロジーとオントロジーマッピングをすることで，時系列データの特徴からオントロジー推論で文章を自動生成する．本システムによりドメイン固有オントロジーを用いて時系列データの特徴を要約した文章を自動生成することを確認した．

## Automatic Generation of Textual Summaries from Time-Series Data using Domain-Specific Ontology

HAYATE KONDO<sup>1</sup> MASAYUKI NUMAO<sup>1</sup>

### 1. はじめに

センサ類の増加や発達により，IoT が普及したことで，非常に多くの時系列データが得られるようになった．時系列データから得られる情報は有用なものが多く，データ分析によって様々な場面で利用されている．しかし，現状では時系列データはそのまま表やグラフで表示されることが多く，専門知識のない人が解釈することは困難な場合であることが多い．そのため，専門知識のない人にも理解できるように，時系列データの特徴を要約した文章を自動生成する研究が行われているが，その多くは株価の変動や気象データの説明文の生成など，対象が限定されるものがほとんどである．本研究では，任意の時系列データから要約文を生成するシステムを提案する．そのためには，対象となる時系列データのドメインに固有な知識が必要であり，それをオントロジーで与える．

本提案の応用として，介護業界における介護日誌の自動

生成について説明する．介護業界では，介護士の業務負担が大きな問題となっており，介護士の業務削減として書類作成や引き継ぎなどの間接業務の削減が最も効率的な業務改善と言われている．そこで，介護士が毎日記録する被介護者の介護日誌をセンサデータから自動生成することで介護士の業務負担の改善を図っている．本研究では，ドメイン固有オントロジーと時系列構造オントロジーをオントロジーマッピングして得られる合成オントロジーと時系列データの特徴からオントロジー推論をすることで文章を自動生成するシステムを提案する．

### 2. 関連研究

#### 2.1 文章生成

時系列データから文章を生成する研究は，ルールベースによる文章生成と機械学習による文章生成の大きく2つの手法に分けることができる．

##### 2.1.1 ルールベースによる文章生成

ルールベースによる文章生成とは，文章を生成するための穴埋めのテンプレートを予め作成しておき，入力データ

<sup>1</sup> 電気通信大学大学院 情報理工学研究所  
情報・ネットワーク工学専攻  
The University of Electro-Communications

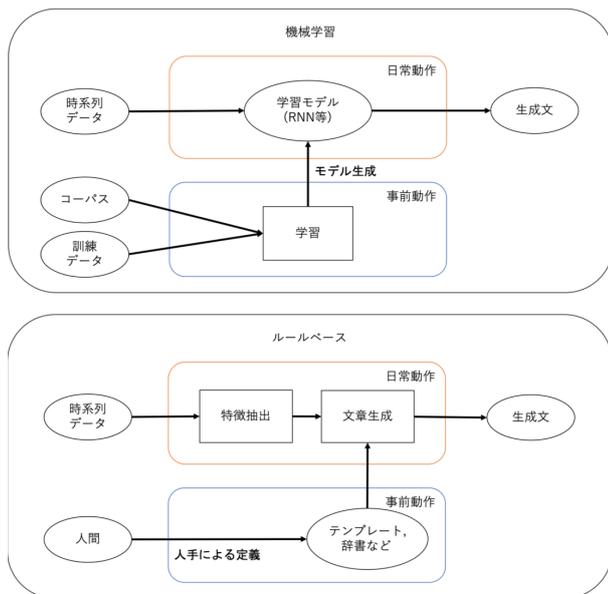


図 1 時系列データからの文章自動生成システムの全体像

によってテンプレートの穴に入れる語句を変更して文章生成をする手法である。テンプレートに沿って文章を生成するため、不自然な文章になりにくいという利点がある。

Ramos-Soto ら [1] は 4 日間の天気予報からそれらを要約した文章を生成するシステムを提案した。この研究では、入力データを中間コードと呼ばれるものに記号化し、予め定義したルールに則って中間コードを言語化し、テンプレートに当てはめることで文章を生成した。入力する天気予報のデータには、天気を示すアイコン (晴れや雨など)、気温、風向きと風速が含まれている。また、ファジィ理論を用いることで、入力データを要約した中間コードの生成を実現させている。提案システムの生成した文章は人が記述した文章に近い自然な文章であることが確認された。

### 2.1.2 機械学習による文章生成

機械学習による文章生成とは、データとそれに対応した文章を訓練データとして学習を行い、その学習モデルにデータを入力することで文章を予測する手法である。

青木ら [2] は株価のデータからその動向を示す文章を生成するシステムを提案した。この研究では、入力データと過去のデータをクラスタリングし、入力データが属するクラスタに分類された過去のデータの動向を示す文書を用いて、文章を生成した。過去のデータの動向を示す文書からバイグラムの言語モデルを構築し、確率的に尤もらしい単語を組み合わせることで、動向内容を示す文章を生成することを確認した。

村上ら [3] は、気象予報モデルをシミュレーションして得られる時系列数値予報マップから文章を生成するシステムを提案した。この研究では、多層パーセプトロン (MLP) を用いて特徴抽出を行うことで数値予報マップから特徴ベクトルを生成し、Long Short-Term Memory (LSTM) を用

いたりリカレントニューラルネットワーク (RNN) を用いて特徴ベクトルから文章を生成した。評価実験の結果、正解テキストに近い内容の文章を生成することが確認された。更に、村上らは同じように RNN を用いて、株価のデータから文章を生成する手法も提案している [4]。

### 2.1.3 既存研究の課題

任意のドメインの時系列データを入力する汎用的なシステムを想定した場合、ルールベースによる文章生成と機械学習による文章生成の 2 つの手法にはそれぞれ課題が存在する。

ルールベースによる文章生成では、入力データのドメインに対応したルールを予め定義しておく必要がある。入力データのドメインによって生成文に使用される語彙や表現は異なるため、予め定義しなければならないルールが膨大な量になると考えられる。さらに、ルールの定義は人手で行うことが殆どであるので、膨大な量のルールを定義することは現実的ではない。

機械学習による文章生成では、入力データと同じドメインの訓練データが大量に必要となる。また、訓練データの内容を表したコーパスも必要となる。従って、任意の時系列データが入力されるため、訓練データとコーパスも入力データのドメインと同じ数だけ必要となる。

## 2.2 オントロジーマッピング

オントロジーマッピングとは 2 つのオントロジーのうち、意味的に同一のクラスを発見することで、2 つのオントロジーを合成する技術である。オントロジーマッピングの手法には大きく分けてクラスの類似性から同一クラスを発見する方法とインスタンスの類似性から同一クラスを発見する方法の 2 つがある。

### 2.2.1 クラスによるオントロジーマッピング

2 つのオントロジーに対して、オントロジーに存在するクラスの名前やクラス構造について類似度を計算し、同一となるクラスをマージすることでオントロジーを合成する。また、名前や構造の類似度だけでは精度が悪かったり意味的に同一のクラスを検出できない場合があるため、外部の知識ベースを利用して同一クラスを決定する手法も存在する。

Mohammed ら [5] はクラスの名前と複数の知識ベースからクラス間の類似度を計算することで、2 つのオントロジーを合成する手法を提案した。外部の知識ベースを用いることで合成対象のオントロジー間だけでは類似度が低いクラスも同一クラスと検出することができる。また、外部の知識ベースを複数利用することで、同一クラスの検出精度を向上させている。

末木ら [6] は日本語版 Wikipedia 記事本文のコンテンツを構造化データとして活用するために、オントロジーマッピングを用いて中間 RDF グラフの生成方法を提案した。

主語、述語、目的語の RDF トリプルに対して、対象の述語と同じ主語と目的語の組を最も多くもつ名前空間上のプロパティにマッピングする。さらに、クラス制約をマッピングに導入することで、主語と目的語のクラスが最頻出する組み合わせを求める。

### 2.2.2 インスタンスによるオントロジーマッピング

2つのオントロジーに対して、オントロジーに存在するクラスに属するインスタンス群の類似度を計算することで、同一クラスを検出しオントロジーを合成する。クラスではなくインスタンスの類似度を計算することで、1つのクラスに対してそれに属する複数のインスタンスを考慮するため、多くの情報を用いて類似度を計算することができる。

Aum Mueller ら [8] は COMA と呼ばれる XML スキーマのマッチングシステムをオントロジーに適用できるシステムに拡張し、COMA++を提案した。COMA++はオントロジーに存在するクラスに属するインスタンスに対してコサイン類似度を計算することで、クラス間の類似度を算出して同一クラスを検出するシステムとなっている。

Duan ら [7] はオントロジーに存在するクラスに属するインスタンス名からなる単語リストを作成し、その単語リストの類似度をクラス間の類似度として利用することで、同一クラスを検出する手法を提案した。また、局所性鋭敏型ハッシュを用いることで計算量などのリソースの問題を解決し、大規模オントロジーのオントロジーマッピングにも適用できる手法を提案した。

## 3. 提案システム

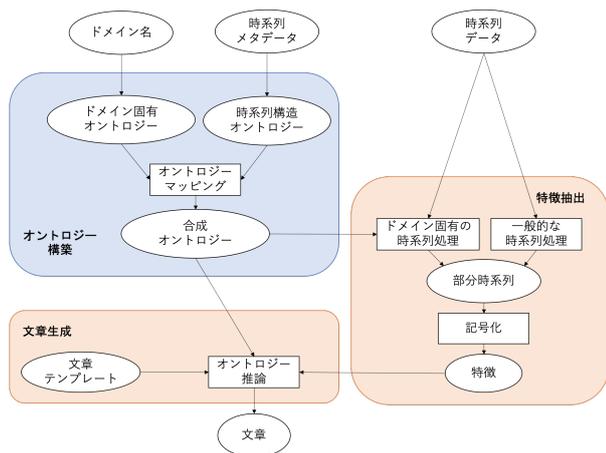


図 2 システム構成図

本研究では、任意の時系列データに対して、特徴を要約した文章の自動生成システムを提案する。本システムはドメイン固有オントロジーと時系列構造オントロジーをオントロジーマッピングにより合成することで、時系列データからドメイン固有の知識を利用してオントロジー推論を行い、文章を自動生成するシステムとなっている。図 2 に示

すように、本システムは時系列データとそのドメイン名、その時系列のメタデータを入力とし、入力した時系列データの特徴を要約した文章を出力する。

提案システムの流れは以下のようになる。

- (1) 事前に、入力する時系列データのドメインに対応するドメイン固有オントロジーと時系列メタデータから時系列構造オントロジーを構築する。
- (2) 時系列構造オントロジーとドメイン固有オントロジーをマッピングし、合成オントロジーを生成する。
- (3) 入力した時系列データに対して時系列処理を行い、その時系列データの特徴的な部分を記号化することで、時系列データの特徴を抽出する。
- (4) 記号化した特徴を用いて合成オントロジーに対して推論を行い、推論結果と文章テンプレートから文章を生成して出力する。

提案システムは大きく以下の3つのモジュールから構成されており、オフラインとオンラインの処理に分けられる。

- オントロジー構築モジュール
- 特徴抽出モジュール
- 文章生成モジュール

次節以降で、各モジュールについて述べる。

### 3.1 オントロジー構築モジュール

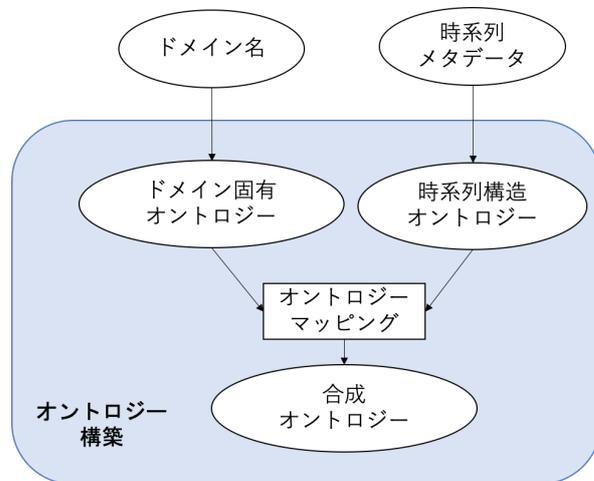


図 3 オントロジー構築モジュール

オントロジー構築モジュールはオフラインの処理であり、システムを動作させる前に使用するオントロジーを構築するモジュールである。図 3 にオントロジー構築モジュールの構成を示す。構築するオントロジーには入力する時系列データのドメインに対応したドメイン固有オントロジーと、入力する時系列データのメタデータから時系列構造オントロジーがあり、構築したドメイン固有オントロジーと時系列構造オントロジーをオントロジーマッピングにより合成することで、最終的な出力である合成オントロジーを

生成する。時系列構造オントロジーは timeseriesML[9] を基に時系列メタデータをオントロジー化したものである。また、ドメイン固有オントロジーは入力する時系列データのドメインごとに人手で構築する。



図 4 ドメイン固有オントロジー

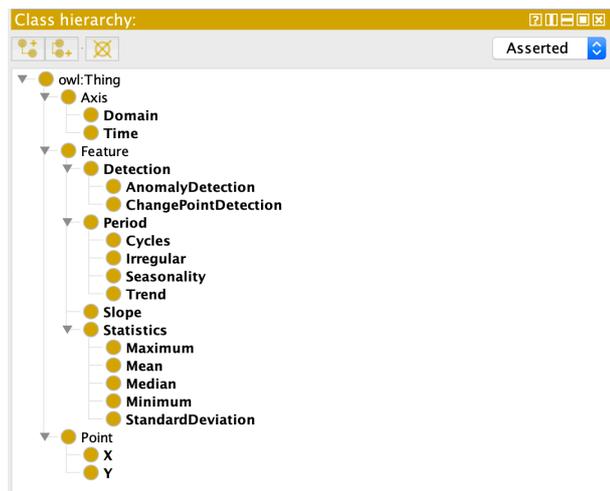


図 5 時系列構造オントロジー

図 4 に構築するドメイン固有オントロジーの例を、図 5 に構築する時系列構造オントロジーを示す。図 4 に示したオントロジーはバイタルサインデータのオントロジーとして、心拍数や呼吸数、血圧などのクラスが VitalSigns クラスの下位に存在し、それぞれに対して、値が通常のもの

高いもの、低いものが子クラスとして定義されている。また、バイタルサインが高かったり低かったりと異常であった場合の症状のクラスが Disease クラスの子クラスとして定義されている。図 5 に示した時系列構造オントロジーは X 軸の時間や Y 軸のドメイン、各点の値のクラスが定義される。また、時系列処理の定義として、傾きや統計、周期や異常検知などが定義される。この 2 つのオントロジーをオントロジーマッピングにより合成することで合成オントロジーを得る。オントロジーマッピングは各オントロジーに含まれるクラスの名前とインスタンスの名前から単語集合を作成し、その単語集合に対してコサイン類似度を計算することで同一クラスを検出して合成する。

### 3.2 特徴抽出モジュール

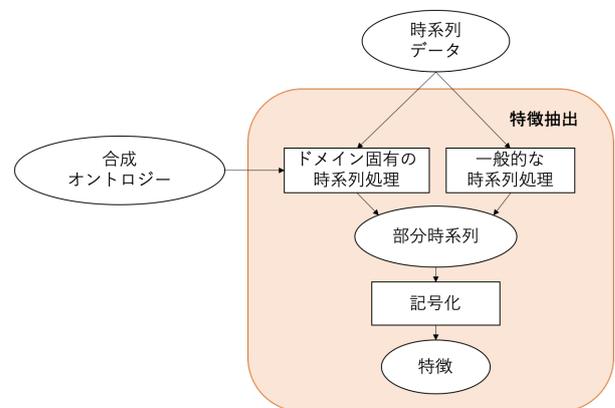


図 6 特徴抽出モジュール

特徴抽出モジュールはオンラインの処理であり、入力した時系列データに対して時系列処理をかけて特徴的な部分を抽出し言語化することで、時系列データの特徴を抽出する。図 6 に特徴抽出モジュールの構成を示す。時系列処理は傾きや変化点、変曲点などを求める一般的な時系列処理とフーリエ変換や異常検知などを使用して周期や部分時系列、季節変動などを求めるドメイン固有の時系列処理に分けることができ、ドメイン固有の時系列処理はオントロジー構築モジュールで生成した合成オントロジーを参照してかける処理を決定する。例として心拍数や呼吸数の時系列データを入力した場合、ドメイン固有オントロジーの異常な心拍数を表す AbnormalHeartRate クラスと時系列構造オントロジーの異常検知を表す AnomalyDetection クラスがオントロジーマッピングによって結びつくため、異常検知を行い特徴と抽出する。最後に、時系列処理の結果を記号化したものを入力した時系列データの特徴として出力する。

### 3.3 文章生成モジュール

文章生成モジュールはオンラインの処理であり、合成オ

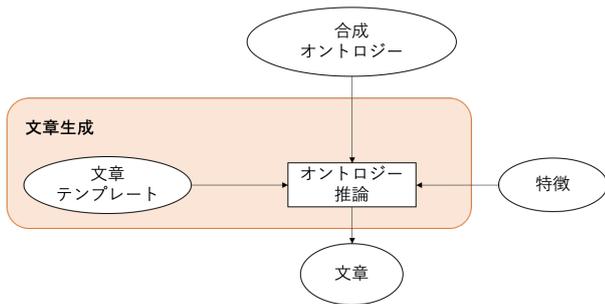


図 7 文章生成モジュール

ントロジーと特徴抽出モジュールの出力である記号化された特徴からオントロジー推論を行う。図 7 に文章生成モジュールの構成を示す。推論した結果を文章テンプレートに当てはめることでドメイン固有の知識を含んだ文章を生成する。例として異常に高い心拍数を検出した場合、記号化した特徴から RapidHeartRate クラスのインスタンスを生成し、オントロジー推論の結果によって Tachycardia クラスが出力される。さらに、文章テンプレートから「頻脈の可能性あります。」といった文章が生成される。

### 3.4 オントロジーの更新

本システムを長期間動作させる場合、周囲の環境に柔軟に対応するための汎用性が求められる。例として介護現場における介護日誌の自動生成を考えた場合、居住者それぞれでバイタルデータの値が変化することはもちろん、同じ居住者でも日々の健康状態は変化する。また、新しいセンサの追加や既存センサの発展により取得されるデータ形式が変化する場合も考えられる。このとき従来のオントロジーをそのまま使用すると、時系列データの処理や言語化が適切に行えなくなる。そのため、既存のドメイン固有オントロジーを適宜アップデートすることが重要である。

居住者の健康状態の変化についてはある一定期間のセンサデータから取得される値を用いてオントロジーをアッ

プデートする。また、入力未知のドメインである場合は Web 検索の結果を用いてオントロジーを拡張する [5]。また、新しい記述様式のドキュメントを作成する必要がある場合には、新しい書式テンプレートを追加して対応できる。

## 4. 評価実験

### 4.1 実験設定

提案した文章生成システムに対して、心拍数と呼吸数からなる時系列データを入力し、バイタルサインのオントロジーを用いて推論を行うことで文章を自動生成することを確認した。

図 8 に入力した心拍数と呼吸数のデータを示す。入力したデータは 11 月 10 日の 0 時から 11 月 11 日の 24 時までの 2 日間のデータであり、上段の時系列データが心拍数データで下段の時系列データが呼吸数である。なお、センサの性質からデータが欠損している時間帯が存在するため補完を行っている。

### 4.2 実験結果

入力した時系列データに対して異常検知とオントロジー推論を行い文章生成を行った結果、以下の文章が出力されたことを確認した。

10 日の 0 時から 2 時まで呼吸数が少なくなる。徐呼吸の可能性あります。11 日の 0 時から 1 時まで心拍数が低くなる。徐脈の可能性あります。

異常検知の結果から異常が発生した時間帯とその異常からどの症状が予想されるかの情報を含んだ文章を自動生成することを確認した。

## 5. おわりに

本論文では、時系列データの特徴を要約した文章の自動

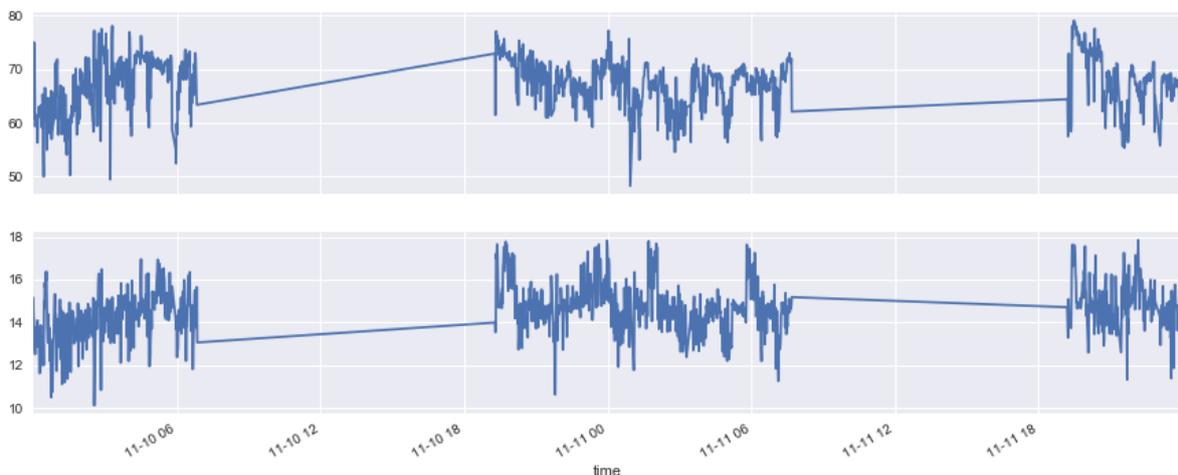


図 8 入力した心拍数と呼吸数の時系列データ

生成システムを提案した。提案したシステムはドメイン固有オントロジーと時系列構造オントロジーに対してオントロジーマッピングで合成したオントロジーを用いて推論を行うことで、ドメイン固有の知識を含んだ文章を自動生成した。また、心拍数と呼吸数の2種類のセンサデータを用いて評価実験を行い、ドメイン固有の知識を含んだ文章が自動生成されることを確認した。

今後の課題としては、オントロジーマッピングや時系列処理などの各モジュールの精度を向上させることで、自動生成される文章の質を向上させることが挙げられる。また、介護業界における介護日誌の自動生成のようにアプリケーションとして応用できるようにシステムの改善を行うことも挙げられる。

## 謝辞

本研究は JSPS 科研費 JP20H04289 の助成を受けたものです。

## 参考文献

- [1] Ramos-Soto A, Bugarín AJ, Barro S, and Taboada J, “Linguistic Descriptions for Automatic Generation of Textual Short-Term Weather Forecasts on Real Prediction Data,” *IEEE Transactions on Fuzzy Systems*, vol. 23, pp.44–57, 2015.
- [2] 青木 花純, 小林 一郎, “時系列データのパターンを考慮した言語モデルに基づく自然言語生成,” *情報処理学会第 78 回全国大会講演論文集*, 595-596, 2016.
- [3] 村上 聡一郎, 笹野 遼平, 高村 大也, 奥村 学, “数値予報マップからの天気予報コメントの自動生成,” *言語処理学会第 23 回年次大会発表論文集*, 2017.
- [4] 村上 聡一郎, 渡邊 亮彦, 宮澤 彬, 五島 圭一, 柳瀬 利彦, 高村 大也, 宮尾 祐介, “時系列数値データからの概況テキストの自動生成,” *言語処理学会第 23 回年次大会発表論文集*, 2017.
- [5] Mohammed Maree, and Mohammed Belkhatir, “Addressing semantic heterogeneity through multiple knowledge base assisted merging of domain-specific ontologies,” *Knowledge-Based Systems*, Vol 73, No.3, pp.199-211, 2015
- [6] 末木 顕人, 兼岩 憲, “Wikipedia 記事からの中間 RDF グラフと DBpedia トリプルの抽出,” *人工知能学会研究会資料*, 45(3), pp.1-7, 2018
- [7] Duan Songyun, Fokoue Achille, Hassanzadeh Oktie, Kementsietsidis Anastasios, Srinivas Kavitha and Ward Michael J, “Instance-based matching of large ontologies using locality-sensitive hashing,” *International Semantic Web Conference*, pp.49-64, 2012
- [8] Aumueller David, Do Hong-Hai, Massmann Sabine and Rahm Erhard, “Schema and ontology matching with COMA++,” *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pp.906-908, 2005
- [9] Lowe Dominic, Taylor Peter, Tomkins James, Cox Simon, Guillaud Frederic, Hershberg Paul, Lindsey Jack, Ritchie Alistair and Utech Michael “A cross-domain standard for representing timeseries data,” *36th Hydrology and Water Resources Symposium: The art and science of water*, pp.995, 2015