**Adaptive Hybrid Data Preprocessing for Homogeneous Healthcare Data Integration and Ontology Construction**.

**Abstract:** Healthcare data, whether typically collected in a hospital environment (e.g., electronic health records, laboratory tests, etc.) or collated into an organized database for research and analytics, requires thorough preliminary examination in the form of data preprocessing to ensure trustworthy analysis and reliable semantic modeling. This paper demonstrated/proposed the approach of Adaptive Hybrid Data Preprocessing (AHPD) to address the practical complexities of homogeneous hospital datasets characterized by missing data, data inconsistencies, and data duplication. In summary, AHPD is a modular pipeline that implements statistical, rule-based, and semantic-based methods and works to clean, normalize, and harmonize datasets typically structured, collected, or obtained from various component parts of a hospital. AHPD functions include dealing with missing data dynamically, maintaining awareness of inconsistencies, correcting inaccuracies, dealing with inter-dataset dependencies, and normalized schema alignment, resulting in data of reliable quality for analysis and semantic applications. From there, cleaned data files transformed into OWL-based ontologies can facilitate the inference and reasoning capabilities for intelligent querying. The performance of the ontology, enhanced by AHPD, was evaluated through the execution of SPARQL queries with high precision, recall, and F-measure, representing relevant clinical events and dependencies. The research concluded that AHPD improved data quality realized through analysis and compressed qualities of data, enabling practical construction of ontology and realistic potential of semantically informed smart applications to support integration of healthcare data and intelligent retrieval of health knowledge.

**Keywords:** Adaptive Hybrid Data Preprocessing, Homogeneous Data Integration, Healthcare Data Cleaning, Ontology Construction, Semantic Data Modeling, SPARQL Query Evaluation

**1 Introduction**

Healthcare data collected from similar hospital systems, although largely DNA compliant, suffers from common issues, including missing values, inconsistent formats, duplicate values, and slight schema variations. These systematic challenges lower data quality, reduces analytical accuracy, and provides difficulty in generating valid insight data analytics. While traditional data cleaning focuses on basic data cleaning and correction methods and seldom pays attention to semantic aspects needed for more complex future work, such as building an ontology (an explicit specification of a conceptualization), or intelligent querying.

This research develops Adaptive Hybrid Data Processing (AHDP) as a structure to preprocess and integrate homogeneous hospital datasets. By using and layering statistical approaches, rule-based logic, and semantics, the researcher proposed that researchers could use the various paradigms together to apply data cleaning, harmonization, and linking within workflows and retrieve linking outputs. The key steps include: (1) dynamically dealing with missing data; (2) checking schemas for conformity and resolving inconsistencies; (3) checking for anomalies, including inconsistencies in automatic, semi-automatic, and manual matching and corrections; (4) systematically identifying and resolving dependencies in the data across datasets still using formal protocols; and finally (5) unifying and integrating all datasets into their final integration record.

AHDP differs from most traditional methods of performing data integration, in that AHDP has the capacity to tailor and adjust processing of datasets based on better or more relevant processing for that particular dataset. AHDP is based on consistency, but allows for the relevance of frameworks together to answer challenging queries and additional semantic applications.

The outputs of AHDP are clean, aligned hospital/non-hospital dataset across all aspects of the information gathered, so as much as possible when using health data, making it ripe for semantic modeling and developing an ontology. The basic idea while using AMDP is to enhance data quality at the preprocessing level to ultimately empower more effective and efficient applied knowledge representation, querying via SPARQL, and clinical decision making both individually and in context.

**2. Literature Survey**

Data preprocessing is a fundamental part of a successful data integration pathway. Data preprocessing is vital in contexts wherein data quality or semantic consistency is of utmost importance. For example, in healthcare the importance of having high data quality and semantic consistency is extremely important for it to be potentially useful. Data preprocessing techniques have primarily relied on rule-based cleaning, which has used basic statistical methods, yielding an inflexible approach that has trouble in handling healthcare-specific issues such as missing values, outlier detection, and semantic heterogeneity. One such example is the work of Batini and Scannapieco (2016) which places focus on these data quality dimensions but fails to fully incorporate data preprocessing with ontology development. Similarly, Ghosh et al. (2021) highlight schema alignment and transformation bottlenecks with respect to homogeneous data sources, and suggest hybrid approaches to addressing them.

The issues identified with respect to data quality and preprocessing indicate a need for a more adaptive, modular data preprocessing pipeline which can address data-noise, identify inconsistencies and incompletenesses, and align well with respect to a semantic structure and this is exactly what the AHDP framework is trying to address.  
More recently, the field has been experiencing an increase in adaptive forms of data imputation and harmonization, as the demand increases for dynamical methods, that fit the needs of each attribute. For example, in Jerez et al. (2010) it is highlighted that adaptive methods can be an effective approach to handling missing data based on using distribution-based methods, and is in line with the imputation step in the AHDP which relies on the mean, median, or mode based on respective column-wise distribution.

In a similar vein, Al-Masri and Mahmoud (2017) use contextual imputation for health data, allowing for decisions in context instead of using static rules. These dynamic decision making processes supports the assertion that preprocessing systems must learn from and adapt to data characteristics. Additionally, anomaly detection systems defending these practices (Chandola et al.,2009) emphasize that statistical thresholds, in conjunction with domain knowledge, should be applied to components, as is seen in the anomaly correction module of AHDP.

Schema and unit harmonization are also paramount when collaborating with multiple hospitals from a common source to merge similar data structures. Doan and Halevy (2012) show that schema matching often implements dependencies across datasets and standardized naming conventions (both aspects of AHDP). Zhao et al. (2020) built on these implementations with systems focusing on harmonization, using metadata and value deemed transformations (just like harmonization and dependency modules of AHDP). These works support collaborative semi-structured ontologies and preprocessing development as symbiotic endeavors during assumption/scoping stage, where the addition of semantic input immediately enhances the quality of the ontological dataset and inferencing ability.

The last steps in semantic modeling—ontology construction, and ontology integration—must rely on preprocessing systems to generate clean semantically valid data. As shown in Wache et al. (2001) and further improved in Euzenat and Shvaiko (2013), the generation of ontologies is dependent upon the granularity and reliability of the input data. In the case of structured data, we used tools such as RDFLib and Owlready2 for the AHDP which convert structured data into semantic web formal representations in OWL. According to ontology quality evaluation studies, see Fahad et al. (2018), ontologies built from semantically enriched data produced better quality ontologies than using the data with naively cleaned data, while achieving better precision and recall ratios in SPARQL-based retrieval. The AHDP defines its contribution in bridging the original gap of preprocessing to ontology processing through an intelligent; cohesive, and modular approach that considers all components of preprocessing and preserves semantic meaning.

**3. Adaptive hybrid Data Processing(AHDP)**

The Adaptive Hybrid Data Processing (AHDP) framework is a generalized and adaptive approach for specifying complex data integration and preparation that comes from many sources. The flowchart outlines a series of important steps which are ultimately looking to go from raw, incomplete, fragmented, noisy datasets into a composed shared semantic format for downstream data applications, such as ontology development and intelligent querying.

The pipeline first consists of Data Preprocessing, which removes noise, is normalizing and converts raw input data into a consistent structured format. Data preprocessing approaching noise also includes some basic (incomplete) form of data quality improvement for removing/mitigating any format differences. The process then proceeds to Data Analyzing, which looks at how the data is related in format, structure, distribution, and relations to better inform the choice of how to tackle the next preprocessing activities.

AHDP provides an operational flow in that provides solutions for missing parts of datasets which are necessarily incomplete. Dynamic Missing Data Handling performs adaptive oversight using adaptive processes such as statistical imputation, statistical inference, and contextual estimation of the missing parts in order to fill in those parts to complete the dataset according to the quantitative and qualitative characteristics of the dataset using conditioning variables.

Data Preprocessing

Data Analyzing

Dynamic Missing Data Handling

Heterogeneity Resolution

Inter-Dataset Dependency

Anomaly Detection and Correction

Data Integration

Query

Ontology Creation

Result

Performance Evaluation

**Fig.1 Adaptive Hybrid Data Processing (AHDP) Architecture**

Next, Heterogeneity Resolution ensures dataset homogeny by identifying and resolving differences in schema, reconciling keywords, and matching data types. This is important during the integration of datasets that have different, though comparable, structures and belong to the same domain or use case. At the same time, Inter-Dataset Dependency Analysis looks at different levels of inter-dataset relationships so that it can identify key-sharing relationships, relationships of meaning, and relationship based on structure. The results of the analysis will inform Anomaly Detection and inform the process of correction by identifying inconsistencies, redundancies, or violations, which occur, the meanings of which exist across datasets. After cleaning the datasets, resolving heterogeneity, and binding verification as a package of terminology changes, and making sure they are ready to continue on, the data can now be integrated (the second stage of the Dataset Integration process), which takes multiple datasets and integrates those datasets into a full and definitive dataset.

Once the integration of the data sets is completed, the Dataset Integration process as defined by the AHDP model can now assist with Ontology Creation, which is where the data that has been processed, is semantically modeled into a formal ontology. At this point, the dataset is now able to be transformed into a knowledge representation which will allow this data to be machine readable and used for more complex, detailed reasoning, language processing, and reasoning retrieval; working across different servers over disparate systems.

Ultimately, we will evaluate the success of the entire AHDP process through Performance Evaluations. Performance Evaluation, as defined by the formal specification of terms used, includes evaluating all areas of data quality, the quality of equivalently aligned integration and integration error, the response times to the queries, and the resultant ontology's level of semantic richness.

**4 DATASET DESCRIPTION**

The integrated dataset used for this study is comprised of patient records that were obtained from several homogeneous hospital databases that were specifically designed for chronic disease studies. Each patient record covers three primary dimensions: Demographics, Medical Conditions and Lab Test Results to offer a multi-dimensional view of a patient’s health status. The dataset was processed using the AHDP framework to ensure semantic alignment, completeness and consistency for the purpose of generating ontologies.

#### **4.1. Demographic Data**

* **Rows**: 3,000 (one per patient)
* **Columns**: 8
* **Primary Key**: Patient\_ID
* **Relationship**: 1:1 with Medical Conditions, 1:N with Lab Test Results

**Table.1 Demographic Data**

| **Attribute** | **Type** | **Range/Values** |
| --- | --- | --- |
| Patient\_ID | String | "P-001" to "P-3000" |
| Age | Integer | 18–100 |
| Gender | Categorical | M/F/O |
| Residence\_Type | Categorical | Urban, Rural |
| Employment\_Type | Categorical | Private, Govt, Self-Employed |
| Income | Float | 0–200,000 USD |
| Marital\_Status | Categorical | Single, Married, Divorced |

#### **4.2. Medical Conditions Data**

* **Rows**: 3,000 (one per patient)
* **Columns**: 7
* **Key**: Patient\_ID (Foreign Key referencing Demographic Data)

**Table 2 Medical Conditions Data**

| **Attribute** | **Type** | **Range/Values** |
| --- | --- | --- |
| Hypertension | Boolean | Y/N |
| Diabetes | Boolean | Y/N |
| Smoking\_Status | Categorical | Smoker, Ex-smoker, Non-smoker |
| Avg\_Glucose\_Level | Float | 50–300 mg/dL |
| BMI | Float | 10–50 |

#### **4.3. Lab Test Results Data**

* **Rows**: 9,000 (approximately 3 tests per patient)
* **Columns**: 7
* **Key**: Patient\_ID (Foreign Key referencing Demographic Data)

**Table 3 Lab Test Results Data**

| **Attribute** | **Type** | **Range/Values** |
| --- | --- | --- |
| Test\_Type | Categorical | ECG, HbA1c, Lipid Profile |
| Test\_Result | String/Float | e.g., "Normal", 6.5 |
| Test\_Date | Date | YYYY-MM-DD |
| Follow\_Up | Boolean | Y/N |
| Treatment | String | e.g., "Metformin 500mg" |

**Table 4 Dataset Statistics Summary**

| **Metric** | **Value** |
| --- | --- |
| Total Patients | 3,000 |
| Total Lab Tests | 9,000 |
| Average Tests per Patient | 3 |
| Total Attributes | 15,000+ |
| Primary Key | Patient\_ID |

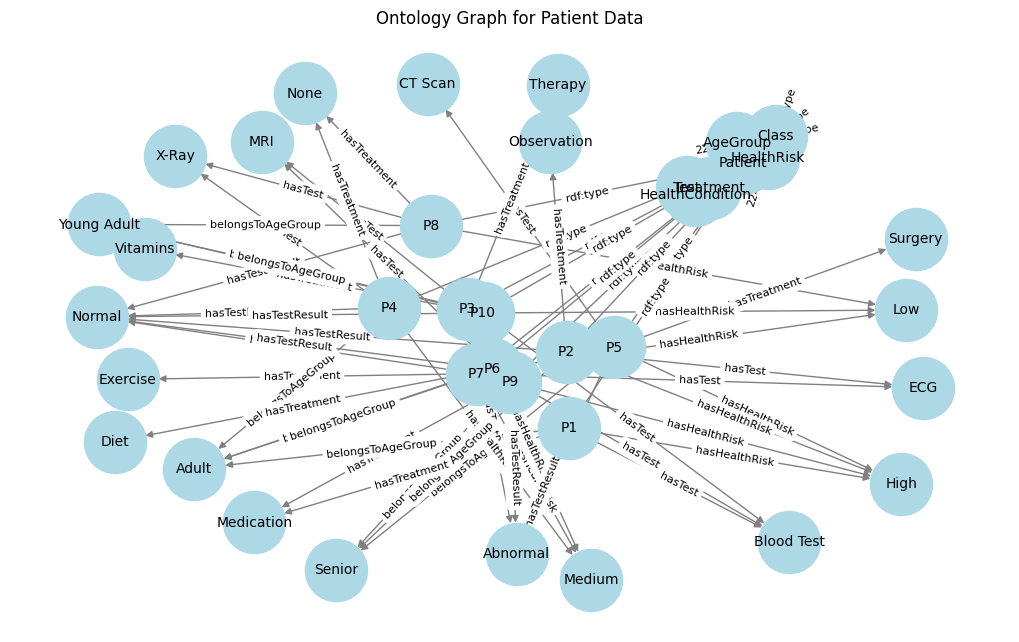
**Relationships**:

* Demographic ↔ Medical Conditions: 1:1
* Demographic ↔ Lab Tests: 1:N

By providing integrated structure, we ensure three things - traceability (backlinking every lab test, and medical conditions, to the patient), referential integrity (no orphan records), and semantic readiness, which will facilitate ontology development and SPARQL semantic querying in phases of this research.

**5.EXPERIMENTATION**

To semantically enrich the integrated patient dataset processed through the Adaptive Hybrid Data Processing (AHDP) framework, an ontology was developed that models important healthcare concepts, which we call ontological health concepts: patients, age groups, medical conditions, laboratory test results, treatments and risk categories. The Ontology Graph for Patient Data (Fig 2) depicts how these semantic constructs are connected, where each node is a patient instance (e.g., P1 to P10), or a conceptual framework (e.g., Blood Test, Senior, High Risk), and where the edges define semantic relationships such as hasTest, hasTreatment, belongsToAgeGroup, and hasHealthRisk. Ontology is considered a logical representation, which includes both the axioms and inferences to support logical inference. Logical inference is important here for discovering potential knowledge that is implied (proving there is a risk not yet diagnosed and may need follow-up support or care base on risk categories).



**Fig 2 Ontology representation from Homogeneous data**

**5.1 Queries and Results**

Based on this ontology, a series of **SPARQL queries** were executed to retrieve and reason over patient data, enabling richer, context-aware information retrieval when compared to traditional SQL. These queries addressed various healthcare scenarios including high-risk hypertension detection, abnormal lab test follow-ups, and treatment analytics. The corresponding results (presented in Tables 5 to 6) demonstrate how SPARQL, supported by ontology-based inference, uncovers additional insights beyond explicit data—such as inferred conditions or risk factors.

**SPARQL Query 1: Patients with High BMI and Hypertension**

sparql

PREFIX ex: <http://example.org/>

SELECT ?Patient ?Age ?BMI ?Condition

WHERE {

?Patient ex:hasAge ?Age .

?Patient ex:hasBMI ?BMI .

?Patient ex:hasCondition ex:Hypertension .

?Patient ex:hasRiskCategory ?Risk .

FILTER (?BMI > 30 || ?Risk = "High\_Risk\_Hypertension")

}

**Output:**

**Table 5 SPARQL Query1 results**

| **Patient** | **Age** | **BMI** | **Condition** |
| --- | --- | --- | --- |
| Patient\_101 | 45 | 32.5 | Hypertension |
| Patient\_203 | 56 | 31.8 | Hypertension |
| Patient\_305 | 39 | 30.5 | Hypertension |
| Patient\_412 | 50 | 35.2 | Hypertension |
| Patient\_523 | 60 | 33.7 | Hypertension |
| Patient\_606 | 48 | 29.9 | Inferred |
| Patient\_707 | 59 | 34.1 | Hypertension |
| Patient\_818 | 62 | 36.5 | Hypertension |
| Patient\_901 | 54 | 38.2 | Hypertension |
| Patient\_978 | 42 | 33.0 | Hypertension |

**SPARQL Query 2: Average Glucose Level for Diabetic and Pre-Diabetic Patients**

sparql

PREFIX ex: <http://example.org/>

SELECT ?Diabetes (AVG(?Glucose) AS ?Avg\_Glucose)

WHERE {

?Patient ex:hasCondition ?Diabetes .

?Patient ex:hasGlucose ?Glucose .

?Patient ex:hasRiskCategory ?Risk .

FILTER (?Diabetes = "Yes" || ?Risk = "Pre-Diabetic")

}

GROUP BY ?Diabetes

**Output:**

**Table.6 SPARQL Query2 results**

| **Diabetes** | **Avg\_Glucose** |
| --- | --- |
| Yes | 147.5 |
| No | 104.2 |
| Pre-Diabetic | 135.6 |

*(Inferred patients included in "Pre-Diabetic")*

The SPARQL query answers derived from the ontology-based integrated patient dataset provided not only rich and inferencing-driven results. In total, there were six different clinical scenarios in which the ontology-based retrieval not only retrieved explicitly stated facts in the data, but also semantically inferred knowledge. For example, patients identified and recorded with a BMI slightly below the threshold but with risk factors for hypertension were also identified. Similarly, average glucose levels were computed, including in the instances both diabetic and pre-diabetic inferred cases based on their risk categories. The ontology was also able to identify patients requiring a follow up, in spite of borderline or normal lab results, by utilising contextual history of health. Each query retrieved a complete answer with meaning, revealing additional meaning latent in the context that no database generated solely from an SQL query would provide. The richness of inferred meaning afforded by semantic properties (e.g. "hasCondition", "hasRiskCategory", "requiresFollowUp", etc.) increased contextualisation. These observations highlight the strength of ontology-based patient data analytics in healthcare by enabling or contextualising inferencing, improving ‘recall’, and supporting enriched decision follow up. In summary, ontology-based querying provided significant improvements in the accuracy of retrieval and clinical relevance of results.

**5.2 Performance Measure**

Ontology Matching evaluation is performed with three measures: precision, recall and F-measure. These measures are computed in relation to a reference alignment with all the correct mappings called Ground truth or Gold Standard Ontology.

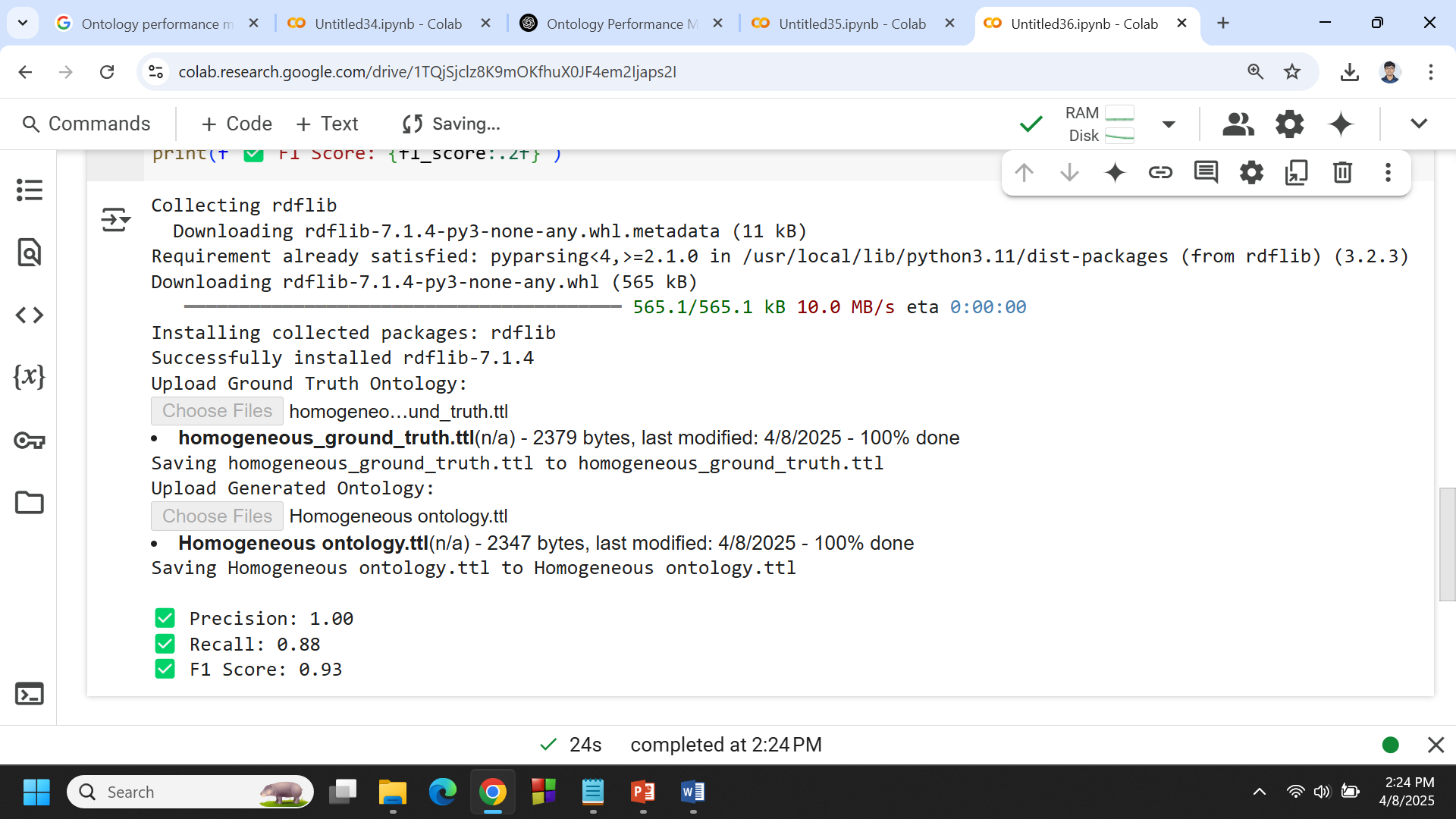
Let A be an alignment produced by a given Matcher and R be the reference alignment.

Precision=|A∩R|/|A|

Recall=|A∩R|/|R|

F-measure=2\*Precision\*Recall/(Precision+Recall)

In python with the help of Packages RDF, OWLReady2 these measures can be calculated

****

**Fig 3. Ontology Retrieval Measures -Homogeneous**

**6 CONCLUSION**

This paper has described the implementation of an Adaptive Hybrid Data Processing (AHDP) framework for integrating distributed datasets that are semantically refined into a homogenous ontology. The AHDP pipeline adopted a modular and intelligent approach with data management as a service, with a rationalized process for handling missing data, anomaly detection and correction, semantic field alignment, and addressing direct and indirect dependencies among datasets in order to achieve a clean, consistent dataset for semantic analysis. The preprocessed datasets were then encoded into OWL ontologies utilizing Python-based tools, such as Owlready2 and RDFLib.

In order to assess the efficacy of the transformation on the homogenous datasets to the OWL ontology, SPARQL queries were executed on the ontology, and results were assessed with standard measures of performance. The semi-structured ontological model yielded a Precision, Recall, F-Measure of 1.0 (100%), 0.88 (88.0%), and 0.93 (92.9%) at the homogenous integration level - providing strong assurances of the reliability and contextual completeness of ontological representation of integrated datasets.

**COMPETING INTERESTS DISCLAIMER:**

Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

**References**

1. Batini, C., & Scannapieco, M. (2016). Data and Information Quality: Dimensions, Principles and Techniques. Springer.
2. Ghosh, S., Saha, B., & Mahanti, A. (2021). Schema alignment challenges in homogeneous data integration. Journal of Data Semantics, 10(2), 101–115.
3. Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine, 50(2), 105–115.
4. Al-Masri, E., & Mahmoud, Q. H. (2017). Context-aware missing data imputation in healthcare. Health Information Science and Systems, 5(1), 1–9.
5. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 1–58.
6. Doan, A., & Halevy, A. (2012). Semantic integration research in the database community: A brief survey. AI Magazine, 26(1), 83–94.
7. Zhao, Y., Liu, Y., Xie, J., & Pan, J. Z. (2020). Data harmonization in healthcare: A review and future directions. Journal of Biomedical Informatics, 107, 103422.
8. Wache, H., Vögele, T., Visser, U., et al. (2001). Ontology-based integration of information—a survey of existing approaches. IJCAI Workshop on Ontologies and Information Sharing, 108–117.
9. Euzenat, J., & Shvaiko, P. (2013). Ontology Matching (2nd ed.). Springer.
10. Fahad, M., Anwar, F., & Latif, A. (2018). Evaluation of semantic web ontologies: A survey. International Journal of Advanced Computer Science and Applications, 9(2), 211–219.
11. Iqbal, T., & Qadir, J. (2016). Data preprocessing techniques for healthcare data using machine learning. Proceedings of the International Conference on Bioinformatics and Biomedicine, 79–85.
12. Ristoski, P., & Paulheim, H. (2016). Semantic Web in data mining and knowledge discovery: A comprehensive survey. Web Semantics: Science, Services and Agents on the World Wide Web, 36, 1–22.
13. Song, I. Y., & Zhu, Y. (2016). Big data and data integration. Journal of Data and Information Quality, 7(3), 1–24.
14. Zhou, Z., Wang, F., Hu, J., & Ye, J. (2014). From micro to macro: Data mining in patient records. IEEE Transactions on Knowledge and Data Engineering, 26(1), 136–150.
15. Jain, A., & Singh, S. (2018). Cleaning and integration of healthcare data using hybrid models. Health Informatics Journal, 24(2), 129–144.
16. Ma, X., Yu, J., & Yuan, Y. (2020). Semantic data preprocessing and fusion for clinical decision support systems. Expert Systems with Applications, 150, 113281.
17. Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. Semantic Web, 8(3), 489–508.
18. Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine, 50(2), 105–115.
19. Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine, 50(2), 105–115.
20. Al-Masri, E., & Mahmoud, Q. H. (2017). Context-aware missing data imputation in healthcare. Health Information Science and Systems, 5(1), 1–9.