Environmental Sustainability (2018) 1:367–381
https://doi.org/10.1007/s42398-018-00033-1

ORIGINAL ARTICLE

Interoperable framework for improving data quality using semantic approach: use case on biodiversity

Priyanka Singh[1,2] · Dheeraj Kumar[2] · Sameer Saran[1]

Received: 26 September 2018 / Revised: 11 November 2018 / Accepted: 13 November 2018 / Published online: 22 November 2018
© Society for Environmental Sustainability 2018

Abstract

Today the Internet growing exponentially and revolutionizing everything with increasing number of users everywhere in order to meet the superfluous demand has triggered an unprecedented wave of various kinds of digital data on the Web. Among them much of the data is relevant and can be turned into actionable insights but difficulties to face are that handling such a hype of data on the Web and due to its unstructured format can not meet the pre-set requirements of professionals and end users. In the context of biodiversity domain, a conceptual approach of data science has been proposed in this paper to extract and structure data seamlessly, which makes sense of all biodiversity-rich data and multiple-record documents by saving time and energy. The major drawback in manual extraction and storage of biodiversity data is that it gives rise to several errors (such as spelling errors, skipping of some data fields etc.) which can be difficult to improve during the processing stage, thereafter can not meet the research demands. However, such drawbacks can be dealt if data science approach is applied within the system and this automated approach will be fast, flexible, reliable and accurate. Nevertheless, the only thing to be taken care in the extraction approach is regular monitoring and analysis of Hypertext Markup Language (HTML) structure, documents, and links of target sources. Such a huge set of data contains many error and noisy characters; to eliminate these errors, data cleaning algorithm has been used to make data error-free and ready for further systematic research. Due to the wide variety of data formats, achieving interoperability is a daunting task, since some of the datasets do not follow their own schema structure. To cope with this demand, semantic interoperability has proved to be helpful by exchanging data through web services between different independent loosely coupled systems. This paper presents an overview of semantic interoperability and case studies on various projects that implemented it for biodiversity data sharing.

Keywords Data extraction · Data mapping · Data cleaning · Biodiversity · Semantic interoperability

Introduction

Each and everything which is easily available and natural has been either destroyed or commoditized and due to this our problems have mounted and mankind is surrounded by the threats which they have themselves invited, such as climate change, loss of biodiversity, changes in biogeochemical cycles, etc. (Arora 2018a). As it is known that necessity is the mother of inventions and humans have time and again come up with ingenious discoveries and there have been noteworthy findings related to environment. For this, digital world has played a vital role in collecting accurate data and establishing a communication with the findings or research by saving costs and efforts associated with paper-based data entry (Silvertown 2009). Nowadays, research on biodiversity is highly dependent on various web-based resources tools and approaches such as online data repositories, web-based data analytics, scientific catalogs, text mining system. Another form of data source is the contribution made by the experts on the web which includes a wide variety of biodiversity data to be managed and analysed. Some examples of web-based tools implemented in biodiversity domain are Integrated Taxonomic Information System (ITIS) that handles taxonomic information on plants, animals, fungi and microbes of North America and World (https://www.itis.

✉ Priyanka Singh
pri2390@gmail.com

[1] Geoinformatics Department, Indian Institute of Remote Sensing, Indian Space Research Organisation, #4, Kalidas Road, Dehradun 248001, India

[2] Mining Engineering Department, Indian Institute of Technology (Indian School of Mines), Dhanbad, Jharkhand 826004, India

gov); Global Biodiversity Information Facility (GBIF). A well-known global repository of bioresources with a strong focus on dissemination of scientific data in the form of web services (https://www.gbif.org); BioVel, an R—based (a programming language) virtual lab for biodiversity and ecology analytics, and modeling through the curated web services (Alex et al. 2016; Hardisty et al. 2016); SuperGIS Biodiversity Analyst, an spatial analytic web tool to assess the richness, evenness and diversities of biodiversity (https://www.supergeotek.com/index.php/products_biodiversity_analyst); Biodiversity Heritage Library, an open catalog of biodiversity literature (Page 2011); the COPIUS project, a knowledge repository of Philippines based biodiversity data, built by the University of Manchester's National Centre for Text Mining big data analytical tools (Batista- Navarro et al. 2017); Global Fish Watch, an online tool to analyze the fishing patterns and safeguarding against overfishing (http://globalfishingwatch.org/map/) and Biodiversity A-Z, a simple tool to help in gaining knowledge about biodiversity conservation (http://www.biodiversitya-z.org). In addition, many mobile applications have been designed and developed with a vision to facilitate the portable and ease contribution by experts, professional and general public members in biodiversity sector. Few examples of biodiversity-related mobile apps are ZSL Instant Wild, developed by Zoological Survey of London (ZSL) (https://www.zsl.org/conservation/conservation-initiatives/conservation-technology/instant-wild) and IBIN mobile app developed for the Indian Bioresource Information Network (IBIN) project using crowdsourcing approach (Singh et al. 2018). In the past few years, an increasing wealth of data in biodiversity sector over the World Wide Web (WWW) has opened many entrants to receive a beneficial opportunity to the end users in many exciting ways (Brin et al. 1998). Generally, users retrieve and access data by browsing the web pages and searching the keyword across the web, this is an intuitive process but leads to several limitations (Apers 1995). Furthermore, browsing for a particular item or keyword in WWW is to find a key in the ocean of vast information which results with a lot of links and following them is a tedious task with the possibilities to get lost in it. If web browsing is compared to keyword search, later is little bit more efficient but often returns a vast amount of information or data, which is difficult for the novice as well as professional users to use efficiently. Like traditional databases, web data can rarely be queried and manipulated, many biodiversity-related tools rely on managing and exploiting this platform of web resources within a stipulated timeframe.

In order to retrieve and handle data from WWW more efficiently and effectively, some researchers have turned to ideas of database techniques (Florescu et al. 1998). However, database techniques require a structured way of data but most of the web data is in unstructured form and prone to colloquialisms, grammatical errors, accents and mispronunciation, therefore, database techniques can not be directly applied to them. To address this issue of handling large scale of web data, a suggested strategy is to implement a data science that includes the extraction of meaningful data from web resources, cleaning of data to make it useful so that it can be used for later processing, analyzing and querying by end users. This extraction process of data from the web sources is known as web data extraction or more, generally known as web scraping and web harvesting.

In this paper, a novel method of data science has been proposed to extract, clean and store data in context of biodiversity in a structured way and tested on few of the global biodiversity websites. After extraction procedure, extracted data will be integrated into the Indian Bioresource Information Network (IBIN) database for data enrichment and will be used as a knowledge base in semantic interoperability model.

Indian bio-resource information network (IBIN) project

According to Singh et al. (2018), our planet is full of unseen species and before their identification, they are getting extinct each year, due to such a heavy loss of bioresources, it is difficult for professionals to plan the conservation and modeling of the geographical distribution of species and habitat analysis. In order to create inventory, analyze, prospect and conserve the bioresources, a number of organizations are working towards generating large datasets. Unfortunately, all such data are highly scattered and not easily accessible with limited possibilities to add value to each other. In context of creating a vast repository of species and make it available through a single web platform, Indian Bioresource Information Network (IBIN), a national project for creating and maintaining a digitized collection of the biological resources of India extracted from the specimens and published literature through a single digital window, was conceived by the Department of Biotechnology, Government of India, in 2006. The program began as a collaborative effort between the Department of Space, (represented by IIRS & NRSC) and University of Agricultural Sciences (UASB). Its major goal was to network and promote an open-ended, co-evolutionary growth among all the digital databases related to biological resources of the country and to add value to the databases by integration. IBIN functions as a common platform where the compiled information on various economic and medicinal plants, animals and microbial resources (from across India) are served to a diverse range of end users through the IBIN portal (www.ibin.gov.in) in the following formats—Spatial and Non-spatial datasets (Roy et al. 2012). Spatial datasets contain maps of

biological richness, vegetation type, fragmentation and disturbance index at 1 km × 1 km grid and non-spatial datasets contains information on medicinally and economically valuable plants, animal, marine and microbial assets of the country (Saran et al. 2012). IBIN portal is designed and developed in order to serve relevant information on bioresources of the country to the professionals involved in bioprospecting, marketing, protecting bio-piracy and conservation.

Historical context

In 1989, when Tim Berners-Lee invented WWW (Berners-Lee 1999), the global content of information was not massive and that era did not need any extraction system to retrieve and scrape data off them. But the recent development and expansion of internetwork of million computers, the need for the data extraction is becoming essential. Earlier, the process of extracting data from web sources was a manual approach and time consuming, which was not sufficient and could not be used on large scale business applications. There were various libraries and software such as spreadsheet and way back machine (Murphy et al. 2007) for simple and small-scale web data extraction tasks. In order to make data extraction tasks viable for large-scale business applications, web scraping services came into the picture and recognized as a preferred route by organizations and research groups, who are seeking for web data. For advanced web extraction, Artificial Intelligence (AI) has been incorporated in web extraction to intelligently identify and extract fields from webpages (Lage et al. 2004). This new research is a stage towards developing an intelligent web spider that can go through the variety of resources like humans, rather than just being engaged in one process. In recent decades of WWW, development was done in two particular ways: searching for content on web pages and following the links interlinked with those particular information (Arocena and Mendelzon 1999). Such development in the field of WWW resulted in an advanced and fast querying technology from the manual libraries on the web to deal with a far digital end.

Related work

This section attempts to present a brief survey on web data extraction tools developed by various research groups. Laender classified the data extraction tools based on the techniques used by each tool to generate wrappers and grouped them as follows: ontology-based tools, modeling based tools, Hypertext Markup Language (HTML) aware tools, wrapper induction tools and Natural Language Processing (NLP) based tools (Laender et al. 2002a, b). Ontology-based tools were developed by Brigham Young University's Data Extraction Research Group, which executes extraction model on the ontologies of specific domain including relationships, lexical appearance and context keywords (Embley et al. 1999). Few examples of ontology-based tools are SUMO (Niles and Pease 2001), DOLCE (Gangemi et al. 2003), On-to-knowledge (Fensel et al. 2000) and Protégé (Gennari et al. 2003). The modeling based tools locate the given target of interest in webpages portions of data, identifies the objects and extract their data. Tools based on this approach are North Western Document Structure Extractor (NoDoSE) (Adelberg 1998) and Data Extraction By Example (DEByE) (Laender et al. 2002a, b; Ribiero-Neto et al. 1999). The HTML aware tools job is to parse the structure of HTML document before the extraction procedure begins, for e.g., RoadRunner (Crescenzi et al. 2001), WysiWyg Web Wrapper Factory (W4F) (Sahuguet and Azavant 2001) and Lixto (Baumgartner et al. 2001). The wrapper induction tools generate extraction rules for a set of training sets or examples, for e.g., WIEN (Kushmerick 2000), SoftMealy (Hsu 1998) and STALKER (Muslea 2001). NLP-based tools use natural programming language techniques such as filtering, part-of-speech tagging, and lexical semantic analysis to build relationships and obtain extraction rules. Some examples of NLP-based tools are Robust Automated Production of Information Extraction Rules (RAPIER) (Califf and Mooney 1999), SRV (Freitag 2000) and WHISK (Soderland 1999). Another survey on web data extraction was carried out by Chang and Lui (2001) and Kayed and Chang (2010) on the basis of the following three dimensions—task difficulties, techniques used and degree of automation to compare different data extraction systems.

The information extraction systems are classified into four categories: supervised, semi-supervised, unsupervised and manually constructed information extraction system (Kadam and Pakle 2014). The supervised information extraction system extracts data from a set of labeled web pages and outputs a wrapper. For example—RAPIER (Califf and Mooney 1999), SRV (Freitag 2000), WIEN (Kushmerick 2000), STALKER (Muslea 2001), SoftMealy (Hsu 1998), NoDoSE (Adelberg 1998), DEByE (Laender et al. 2002a, b; Ribiero-Neto et al. 1999) and WHISK (Soderland 1999). The semi-supervised information extraction system uses user's examples to generate rules and their examples are OLERA (Chang and Kuo 2004), Thresher (Hogue and Karger 2005) and IEPAD (Chang and Lui 2001). The unsupervised system does not have any user interactions and any labeled training sets. Some of the examples of this category are RoadRunner (Crescenzi et al. 2001), Data Extraction and Label Assignment (DELA) (Wang and Lochovsky 2003) for web databases, Data Extraction Based on Partial Tree Alignment (DEPTA) (Zhai and Liu 2005) and EXALG (Arasu et al. 2003). The manually constructed information systems are developed by the general purpose programming

languages, for e.g., TSIMMIS (Hammer et al. 1997), WebOQL (Arocena and Mendelzon 1999), MINERVA (Crescenzi and Mecca 1998), XWRAP (Liu et al. 2000) and W4F (Soderland 1999).

Proposed methodological framework

Despite the fact that web data extraction has been continuing since a long while, it is rarely being used in the vast field of biodiversity. According to the International Union for Conservation of Nature (IUCN) report on the number of threatened species, the estimated number of described species from 1996 to 2018 is approximately 17 lakhs and only 93,577 species are evaluated till now. This report reveals that only less than 80% of the known species are described due to insufficient coverage (IUCN Red List 2018).

This paper proposes a methodology that includes various steps such as web data extraction, data cleaning, data mapping, data integration, and storage of web data to create and curate a knowledge base for data enrichment, data validation and semantic interoperability tasks, flowchart is shown in Fig. 1 with all the research steps in Fig. 2.

Web data extraction

In the repository of Global Biodiversity Information Facility (GBIF) and other biodiversity databases, there are more than 435 million records of biodiversity data available through the portals, but still, they are not enough to complete the coverage of global biodiversity (Ballesteros et al. 2013; Yesson et al. 2007). Traditionally, data was collected from museums, scientific literature, and specimens to create a knowledge base, but new data sources such as Internet should also be explored so that the manual task of data collection can be eliminated (Blagoderov et al. 2012). The process of extracting data from web sources is known as web data extraction. Web data extraction is a type of software application or process that interacts with and exposes the resources available on the web for the extraction (Laender et al. 2002a, b; Baumgartner et al. 2009) that can be processed, cleaned, mapped and stored for later usage (Irmak and Suel 2006; Chidlovskii et al. 2000). Web data extraction software is widely used in various applications because of its efficiency to collect a large amount of data or information with the limited manual effort. The following techniques are studied in this paper which are used for web data extraction:

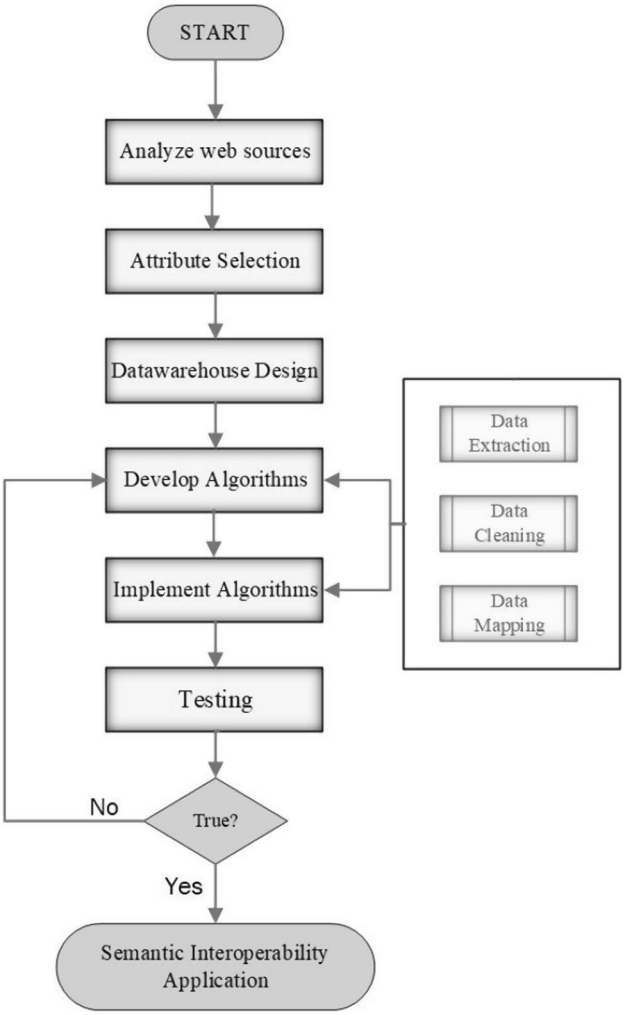

Fig. 2 Flowchart of working web data extraction process

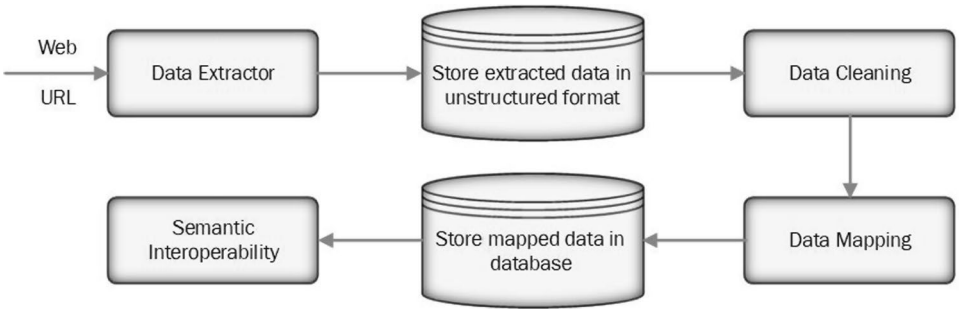

Fig. 1 Framework of proposed approach of web data extraction

a. Tree-based approaches;

   i. Partial Tree Alignment: This algorithm relies on the extraction of those data records collected in contiguous regions of the webpage and can be aligned with certainty (Zhai and Liu 2005, 2006). First of all, it splits webpages into segments based on visual information to identify the gaps between data records on the webpage. After this, the extraction of each data record from its Document Object Tree (DOM) sub-tree position, the data is aligned and stored in the database in a structured way. However, the drawback of this technique is that it is time-consuming in case of complex HTML document structures.
   ii. Tree edit distance matching: This algorithm computes a tree edit distance between the trees and performs insertion, deletion, and replacement of nodes. The first web data extraction approach is based on the tree edit distance matching algorithm developed by Reis et al. (2004). This approach relies on a different type of mapping called Restricted Top-Down Mapping (RTDM) in which all the three sets of operations are limited to the leaves of trees and Yang's algorithm is applied to find RTDM (Yang 1991).
   iii. Simple tree matching: This is considered as an effective solution to overcome the limitations of the tree edit distance matching algorithm (Selkow 1977) by evaluating the similarity between two trees. The only disadvantage of this is that the permutation of nodes cannot be matched and no level crossing is allowed (Ferrara 2014).

Therefore, tree-based approaches are broadly used as they are easy to implement and handle.

b. Regular expression-based approach: This approach uses powerful formal language to generate data extraction rules from webpages that identify the patterns in unstructured text based on some string matching criteria.
c. Machine learning-based approach: This approach fits to extract domain-specific information from manually labeled webpages with different structures.

Following are the software used in web data extraction system

a. Libraries—this is one of the simplest approach which allow developers to build the applications using those programming languages in which they are well-acquainted. These libraries first establish a GET request to the Hyper Text Transfer protocol (HTTP) server to grant an access to the server and inspects DOM structure of webpage to identify the target data nodes. Then, the extracted contents are interpreted and compiled using various functions such as tokenization, matching of regular expression and trimming, and node processors are designed for output data in the user-defined format. Different libraries are available in different programming languages. Python is one of the most popular open source programming languages widely used for web mining, machine learning, natural language processing, network analysis and geospatial analytics because of its ease and a rich ecosystem of available libraries. One of them is BeautifulSoup (https://www.crummy.com/software/BeautifulSoup/bs4/doc/) (Smedt and Daelemans 2012), an HTML parser-based library to extract data by applying different filters to targeted webpages. Similarly, other python based libraries for web data extraction are urllib2, scrapemark and mechanize. In JAVA, jsoup is an open source library for web data extraction, written by Jonathan Hedley in 2009, to parse HTML body from the given Uniform Resource Locator (URL), file or string, extract, manipulate the HTML attributes, clean content to avoid Cross-Site Scripting (XSS) attacks and output the normalized form of extracted web data. Another JAVA library is Jaunt (http://jaunt-api.com/) for extraction, automation and JSON querying. Similarly, Goutte is also an open source scraping library to parse websites and extract data from HTML/XML form of responses. Moreover, there are many libraries written in different programming languages such as rvest for R (https://cran.r-project.org/web/packages/rvest/rvest.pdf), WWW: Mechanize for Perl (https://metacpan.org/pod/WWW::Mechanize) and Nokogiri in Ruby (http://nokogiri.org/). Thus, the libraries provide a flexibility to the developers of using the programming languages as per their own expertise so that they can be aware of the syntax, functions, exceptions and errors.
b. Framework—Libraries have their own limitations, often, several libraries need to be integrated into a single library for the access of webpages and some libraries to parse and extract the contents from web documents. If some changes are done in HTML resources, these libraries get affected as they are modified as per the analyzed view of webpage, and needs recompilation and redeployment of the entire application. Web extraction frameworks provide a solution for the aforementioned drawback of libraries. For instance, scrapy (https://scrapy.org/) is an easy, fast and powerful framework, build in Python, for automated navigation through the websites. The main advantage of using scrapy is that it is built on the asynchronous networking framework, in

other words, users will not have to wait for a finish message or alert before making another request. Similarly, MechanicalSoup (https://mechanicalsoup.readthedocs.io/), PySpider (http://docs.pyspider.org/en/latest/Quickstart/) and Portia (https://portia.readthedocs.io/en/latest/getting-started.html) are other python-based scraping frameworks.

Various web scraping frameworks have been designed and developed in domain-specific programming languages for a particular context of research and therefore, treated as an independent and external artefact. An example of this is WebHarvest (http://web-harvest.sourceforge.net/), a Java based web scraping framework that can be easily implemented and linked to the existing methods. It provides a set of processors as a function and could be combined with a pipeline for data handling and control flow.

Data cleaning

Data cleaning is a method that removes unwanted texts or attributes from the data or information to make them available for analysis and business applications. Data cleaning approach include basic operations such as removing noise and errors, collect necessary information, strategies to handle missing data fields and make a log of known changes (Fayyad et al. 1996). Since it is known that a bunch of data are available on websites but only some of them are useful by using the web extraction approach it is not easy to extract only the useful data because if the extraction model runs on any website, all the data, regardless of whether it meets user's requirements or not, will be extracted. For example—column names and a lot of noisy characters such as new lines, spaces or tabs do not make any sense in the extracted data. Therefore, the data cleaning algorithm has to be executed on the extracted data before making it available for further applications or research purposes.

In this methodology, after the data extraction step, a data cleaning algorithm is applied to clean the extracted data and make it in a valid structured format. This data extraction algorithm has been tested on the extracted data stored in the table format on a webpage, as well as on many data attributes such as table headers, column names, spaces, and line breaks. Therefore, to remove those noisy attributes, data cleaning algorithm has been developed so that the data can be used for further processing tasks.

Data mapping

Data extracted from varieties of web pages is required to be harmonized because websites may follow different and unique naming conventions. Data mapping is a type of cross-referencing approach across different terminologies or naming systems, vocabularies, databases and classification system (Wilson 2007). Mathematically, data mapping is defined in Zhao et al. (2007) as "for any attr∈ ATTR (attr is an attribute and ATTR is a set of all relevant attributes), through mapping rule f, get o = f (attr), we call this a data mapping relation from attr to o where o is a mapping relation entity."

Before beginning of the data mapping process, there are few points to take care of. First of all, different data providers store their data in different types of databases (for e.g. MySQL, PostgreSQL, Oracle, NoSQL, MS Access, MongoDB), and the number and datatype of attributes may vary. Second, source terminology and target terminology can vary, for example, at the source site the name of data is represented as "Name" or "Given Name" and at the target site that particular data is named as "Full Name". Therefore, for the accurate and proper mapping of discrete elements, understanding of how data is represented and how much both the target and source terminologies are similar is required because the mapping process will improve the quality of aggregated data.

For this methodology, data mapping has been done manually by visually analyzing each and every attribute of species web page and mapped with the IBIN database field for successful data integration.

Data integration

Data Integration has been a trendy topic during the past few years with a strong focus on schema mapping (Raghavan and Garcia-Molina 2001). Data Integration is a process of combining different datasets extracted from myriad sources and represents them in a uniform and simplified view to the end users (Halvey 2001; Hull 1997; Ullman 2000). The purpose behind implementing data integration approach is to provide a common format for access within or across the organizations in the realm of information system (Huber 1990; Malone et al. 1987). The key advantage of implementing data integration is that it liberates user from searching relevant sources based on their query, interaction with each data sources autonomously and manually store the raw data from heterogeneous sources. The issues of data integration have received a noteworthy research in various research communities of Artificial Intelligence (AI) and several commercial applications (Chawathe et al. 1994; Stonebraker et al. 1996; Levy et al. 1996; Friedman and Weld 1997; Arens et al. 1996; Cohen 1998; Adali et al. 1996; Woelk et al. 1995; Tomasic et al. 1998; Blakeley 1997; Haas et al. 1997). The changing needs of diverse users have been incorporated into a single integrated data structure through the various means of networking (Bonczek et al. 1978; Date 1995). An instance of such integration is an entity-relationship (ER) model. An ER model is proposed to work the interrelated

data in a conceptual way and provide the basis for integrated information systems (Chen 1976; McCarthy 1982). However, other research groups may perceive that the concerned organizations do not begin from a fresh start and the existing systems are not integrated. In lieu of this, various hypothetical approaches have been developed for the integration of existing data schemas (Batini et al. 1986). Still, others have been focused on the development of practical techniques through which the organizations or research groups could convert their non-integrated systems into the integrated one (Finkelstein 1989; International Business Machines Corporation 1978; Martin and Finkelstein 1989).

An exemplary issue in data integration is that different data providers develop their own integration system to satisfy their own requirements which differ from place to place and resulted in a heterogeneous system with a dependency on software, applications, and databases (Sheth and Larson 1990; Kossmann 2000; Halevy and Ordille 2006). Due to the technological differences in hardware, software and communication system, the different type of heterogeneity occurs (Sheth 1999). This heterogeneity issue leads to the investigation of interoperability in two ways: interoperability of data exchanged and interoperability between heterogeneous systems (Sonsilphong et al. 2012).

Semantic interoperability

For over two decades, interoperability for data and information has been a basic requirement due to the creation, maintenance, utilization and accessibility of heterogeneous sources from the Internet, web and distributed computing infrastructures. Ceccarelli (1997) provided the following definition of interoperability:

> "Interoperability is the ability of two or more software artefacts (e.g. services) to interact effectively at runtime to achieve shared goals, for e.g., a joint activity. The interoperation of two software artefacts A and B requires that A can send requests, R to B based on the mutual agreement and understanding of R by A and B, and B can return responses, S to A based on R by A and B."

After successful data mapping and integration, interoperability is important. Based on the different forms of heterogeneity, interoperability is classified as system, syntax, structure and semantic (Ouksel and Sheth 1999). This section attempts to present a broad survey on semantic interoperability, as shown in Fig. 3, which is being difficult and complex because of its variety of approaches, different requirement levels, number of technical specifications, huge literature and much more examples. The research on interoperability has divided it into three generations:

a. First generation—this generation covers a period of roughly 1980 and their emphasis was to achieve system interoperability for addressing the needs of data exchange within enterprises and departments with different hardware and software, including multi-databases (Liwin et al. 1990) and federated database systems (Heimbigner and McLeod 1985). The believable research in this generation was:

   i. Analysis of structured database schema and innovative approaches to tackle the translation issue with different schema;
   ii. Understanding the schema integration methods (Batini et al. 1986) and level of complexity to perform integration on the real-world schema (Sheth 1995);
   iii. Found ways for schema-based heterogeneity and integration of different heterogonous sources (Drew et al. 1993; Meersman 2005);
   iv. Handling consistencies between different database management systems and advanced

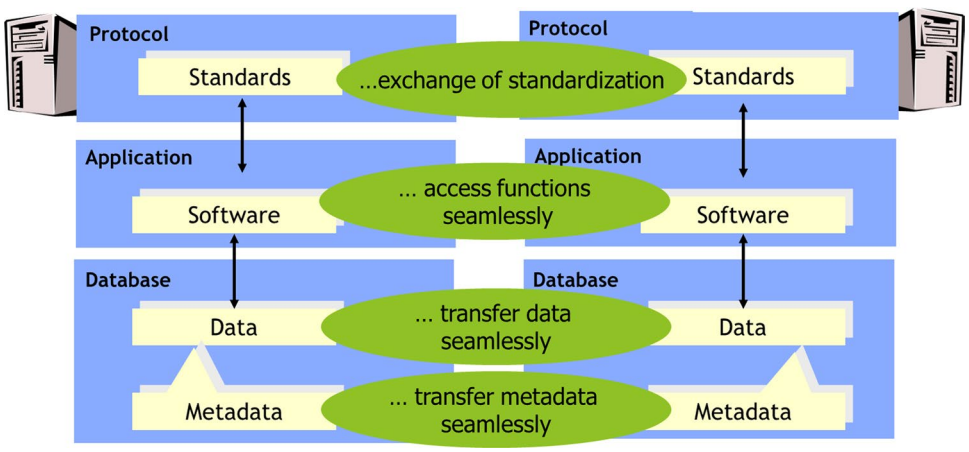

Fig. 3 Interoperability between two systems

research in transaction models (Elmagarmid 1992) and multi-database transaction execution (Georgakopoulos et al. 1994);

The federated database system, first coined by Hammer and McLeod (1979) and later modified by Heimbigner and McLeod (1985), consists of a schema, processors to handle heterogeneity and integration of distributed databases (Sheth and Larson 1990). Based on the management of federation and integration, federated databases are categorized into two broad sub-categories:

i. Loosely coupled system—this is also known as interoperable database system (Litwin and Abdellatif 1986). Such a database system is said to be loosely coupled if the user is having full privileges to create and maintain the federations, and the system administrators have no charge of it. This requires more user engagement and supports dynamic and flexible federated schemas but with less transparency and no updates.
ii. Tightly coupled system—this type of federated database system is said to be tightly coupled if administrators have the full control to create and maintain the federations. This may have one or more stable federated schemas and requires administrator engagement.

b. Second generation—this generation focuses on adopting standardization to achieve system, syntactic and structural interoperability. System interoperability relies on the standardization of Internet communication protocols for the interconnection of two different systems. Suppose there are two systems, S1 and S2, and a vendor V1 implements S1 and vendor V2 implements S2. The model which performs interoperability operation on both, S1 and S2, has complete information of both the systems, mutually shared by the system vendors. Syntactic interoperability deals with the exchange of information based on predefined standards such as HTML and Z39.50. Veltman (2001) described the challenges involved in the syntactic interoperability:

    i. Identification of each and every element in different systems,
    ii. Generating rules to arrange elements in a uniform manner,
    iii. Schema mapping of elements, and
    iv. Mutual agreement and acceptance on rules to bridge different repositories and catalogues.

Structural interoperability refers to the interoperability of data between different information systems having different structures (Wache et al. 2001). This is achieved by the adoption of metadata standards in various domains. Metadata standards are derived from the other global standards of specific domains as per the system needs.

c. Third generation—the rapid development of web and internetworking of global information systems raises an issue on how to cope with the heterogeneous sets of information that consists not only the digital data, but operations and computation to form a new type of data and information. Various approaches or strategies have been implemented that lead to the poor quality of data and information overload. This issue of information overload leads to challenging of "So far (schematically) yet so near (semantically)" (Sheth and Kashyap 1993) turned into "So near (syntactically and structurally) yet so far (semantically)."

Based on the second generation's research, the third-generation results in the formulation of semantic interoperability. Semantic Interoperability implies the semantic mapping or matching of data shared by different data providers, i.e., it does mapping of the same real-world entities representing different meanings in different systems (Nowak et al. 2005).

Shekhar (2004) discussed all the levels of interoperability as follows:

> "The syntactic interoperability specifies common message formats (e.g. tags and marking) to interchange spatial data, patterns and relationships; and the structural interoperability provides means for specifying semantic schemas for sharing; while semantic interoperability involves an agreement about content descriptions of spatial data, patterns, and relationships."

Since the Tim Berners-Lee introduced semantic web in 1999 (Berners-Lee 1999), there has been an increasing demand and interest in W3C standards to provide well-defined semantic exchange and integration of documents and services. It is not concerned with the syntax of data, but a strong focus on the semantics (meaning) of data by adding its metadata and each element of data is linked with the controlled vocabularies. These vocabularies can be the foundation for machine interpretation, knowledge discovery, inferencing and data federation between different information systems. Berners-Lee (1998) said the following on the limitations of the semantic web:

> "A Semantic Web is not Artificial Intelligence. The concept of machine-understandable documents does not imply some magical artificial intelligence which allows machines to comprehend human mumblings. It only indicates a machine's ability to solve a well-defined problem by performing well-defined operations on existing well-defined data. Instead of asking

machines to understand people's language, it involves asking people to make the extra effort."

Though the data are available from different data providers where data provider uses their own standards for data dissemination, integration, and storage, for such cases, ontologies and web services are used for dissemination of platform independent data and also by providing a common shared data encoding schema across disparate parties and as well as establishing mapping rules between local ontology and global ontology. In the context of interoperability, data mapping process can be distinguished into three types, as mentioned by Ghawi and Cullot (2007):

a. Schema Mapping—this type of mapping process proved as fundamental in resolving the number of interoperability issues (Burnstein and Haas 2008) such as schema integration, schema evolution, data exchange, and sharing. For such mapping, two schemas are taken as an input and after applying the schema mapping rules, outputs a mapped schema between the elements of input schema. Examples of schema mapping can be seen in Fuxman et al. (2006) and Miller et al. (2000).
b. Ontology Mapping—this mapping process relates to the vocabulary of two ontologies that share the specific domain. Kalfoglou and Schorlemmer (2003) shows the detailed survey on ontology mapping. The ontology mapping is carried out at between the integrated global ontology and local ontology (Beneventano et al. 2003; Doan et al. 2003; Calvanese et al. 2001), local ontologies (Silva and Rocha 2003) and ontology merge and alignment (Fridman and Musen 2000). Each ontology differs according to their strengths and drawbacks (Choi et al. 2006).
c. Database-to-Ontology Mapping—this mapping process is achieved between databases and ontology by establishing a semantic relation between discrete database elements and ontology components (Ghaw and Cullout 2007).

The term "biodiversity" fathomed the biological diversity at three levels—genetic, species and ecological diversity (Shanmughavel 2007). Genetic diversity alludes to the variation of genes within species. Species diversity refers to the varieties of species. Ecological diversity deals with the ecological processes, biotic communities, and habitats within the ecosystem. Many research groups and organizations attempt to interoperate their biodiversity data by adopting and implementing the globally recognized data standards and interoperability supported framework or architecture. Following are the few examples of semantic interoperability in biodiversity domain:

a. Biological Collection Access Service for Europe (BioCASE)—BioCASE project aims to establish a platform to make specimens and biological data freely accessible by the potential end users through data portals and web services. BioCASE adopted a widely accepted Access to Biological Collection Data (ABCD) standard and eXtensible Markup Language (XML) schema, for the integration and communication purposes. The semantic interoperability approach of BioCASE translates the queries output in ABCD dataset and requests are made in form of XML using HTTP protocol.
b. Species 2000 Interoperability Co-ordination Environment (SPICE 2000)—the goal of this project is to create a complete catalog of all known species data with an assurance of high quality and full coverage (Jones et al. 2000). Such catalogs are coordinated by the taxonomists, researchers, and ecologists of different institutions or research groups (Jones 2006). SPICE 2000 is used by the Global Species Datasets (GSDs) which takes care of data dissemination in the form of responses to the requests made by the users. SPICE 2000 provides searching facility to find synonyms for species name through the joint web services of two Catalogues of Life—Annual Checklist and Dynamic Checklist (Species 2000 Secretariat 2009).

Experimental results

The proposed model is not yet implemented but tested to illustrate the performance, strength, and weakness and this returned a positive output. For the testing purpose, the various species portal has been chosen as a case study for the proposed approach. The first step of the proposed methodology is a web data extraction that returns the scientific names with their associated webpage links of all the species available on the IUCN portal, as shown in Fig. 4. Later, by navigating through the links of species, all the available data of each species are extracted and saved locally in the form of a text file on the local server, as shown in Fig. 5. All the original data are extracted with the column names and a lot of noisy characters, due to this all the extracted data have been passed into the data cleaning method, where they are cleaned and left with the reliable data, as shown in Fig. 6. After this, the mapping process begins on cleaned and formatted data for the attributes such as scientific name, taxonomy, habitat, uses, etc., and all the species records are indexed into the IBIN database. The total execution time taken to extract the complete account of each species is 1 s only, therefore, it can be seen that how the web data extraction changes the manual approach of extraction into the fully automated one.

This model will facilitate to enrich the existing database and make it available as a knowledge base for data quality

```
title: Search Results

URL: /details/3/0
Scientific Name: Aaadonta angaurana

URL: /details/4/0
Scientific Name: Aaadonta constricta

URL: /details/5/0
Scientific Name: Aaadonta fuscozonata

URL: /details/6/0
Scientific Name: Aaadonta irregularis

URL: /details/7/0
Scientific Name: Aaadonta kinlochi

URL: /details/8/0
Scientific Name: Aaadonta pelewana

URL: /details/9/0
Scientific Name: Aaptosyax grypus

URL: /details/61120/0
Scientific Name: Abactochromis labrosus

URL: /details/36563/0
Scientific Name: Abarema abbottii

URL: /details/30843/0
Scientific Name: Abarema bigemina

URL: /details/36553/0
Scientific Name: Abarema callejasii

URL: /details/36568/0
Scientific Name: Abarema centiflora

URL: /details/36560/0
Scientific Name: Abarema cochliocarpos

URL: /details/36549/0
Scientific Name: Abarema commutata

URL: /details/62466/0
Scientific Name: Abarema curvicarpa
```

Fig. 4 Output of data extraction model

check and semantic interoperability with a provision to provide feedback or suggestions on the quality of data. This is an attempt to make a complete information available with full coverage of all the species on a single platform for the end users or professionals involved in biodiversity monitoring, analysis, and conservation tasks.

Discussions

The methodology proposed in this paper is to implement a novel algorithm developed using data science approach in order to extract valuable data and information from varieties of web sources, which are globally recognized in biodiversity domain, cleaning of extracted data to remove noisy characters like whitespaces, tag names, blankspaces, etc., designing schema for mapping and storage of extracted data to the local schema. The major aim behind this research is that no platform ever delivers complete data and information about bioresources, for instance, the IUCN Red List portal (https://www.iucnredlist.org/) delivers comprehensive information on global conservation status of more than 93,500 species along with their geographical distribution; Global Biodiversity Information (GBIF) (https://www.gbif.org/) is an open access portal that provides species occurrences data holding from various institutions around the world in the Darwin Core Archive format and an open source tools to share the occurrences data; and the Indian Medicinal Plants Database (http://www.medicinalplants.in/) documents unique and well-researched information on medicinal uses of plants in Indian systems of medicine. For example, Taverna-based Data Refinement Workflow is designed to integrate the taxonomic data retrieval, data cleaning and geo-temporal data selection to make data fit-for-purpose for a needed research (Mathew et al. 2014). Another work likely similar to this methodology can be seen in O'Sullivan et al. (2010), in which authors capture data related to the plants, invertebrates and birds of forest data in three-dimensional structure to generate the physical structure of forests and can predict six biodiversity measures of the species richness and the abundance of beetles, birds and spiders. Henceforth, a unique, complete and well referenced dynamic repository of biodiversity, linked with all the entities that are distributed on different websites, will be constructed using the algorithm discussed in this paper and will be available for researchers, conservationists, academicians, resource managers, school students, ethnobotanists, and nature enthusiasts. Tragically, biodiversity is the basic requirement of invaluable ecosystem services which is disappearing at the unprecedented rate due to the human civilization because we are lacking in knowledge on biodiversity and its economic and medicinal importance. This loss of knowledge can be prevented by developing an indigenous knowledge base that incorporates various dimensions of biodiversity (Arora 2018b). The knowledge base system will provide a glimpse of what we are having, what is most vulnerable, which area would be fully protected for sustainable use and enrichment of biodiversity, and how to reconcile the increasing demand of society with a link to sustain for biodiversity.

Fig. 5 Unstructured data output

```
Animalia
Mollusca
Gastropoda
Stylommatophora
Endodontidae
Scientific Name:
Aaadonta angaurana Solem, 1976
Red List Category & Criteria:
Critically Endangered (Possibly Extinct) B1ab(iii)+2ab(iii) ver 3.1
Year Published:
2012
Date Assessed:
2011-08-22
Assessor(s):
Rundell, R.J.
Reviewer(s):
Barker, G., Cowie, R., Triantis, K., García, N. & Seddon, M.
Facilitator/Compiler(s):
Pippard, H.
Justification: This is a minute ground-dwelling snail for which no shells
Date last seen:
1936
Previously published Red List assessments:
1996 - Data Deficient (DD) Download assessment 1994 - Vulnerable (V) Downl
Range Description:
The historical geographic range of this species was less than 8 km2 of the
Countries occurrence:
Possibly extinct: Palau
♦ Estimated area of occupancy (AOO) - km2: 0-8 ♦ Number of Locations: 1 ♦
♦ Estimated area of occupancy (AOO) - km2:
0-8
♦ Number of Locations:
1
```

Fig. 6 Cleaned data output

```
Animalia Mollusca Gastropoda Stylommatophora Endodontidae
Aaadonta angaurana Solem, 1976
Critically Endangered (Possibly Extinct) B1ab(iii)+2ab(iii) ver 3.1
2012
2011-08-22
Rundell, R.J.
Barker, G., Cowie, R., Triantis, K., García, N. & Seddon, M.
Pippard, H.
This is a minute ground-dwelling snail for which no shells (live or dead) have been found in recent years,
1936
1996 - Data Deficient (DD) Download assessment 1994 - Vulnerable (V) Download assessment
The historical geographic range of this species was less than 8 km2 of the heavily modified island of Ang
Possibly extinct: Palau
Estimated area of occupancy (AOO) - km2: 0-8, Number of Locations: 1,  Lower elevation limit (metres): 1,
1
1
```

Conclusions

In such cases when a large amount of data is available but still the fields are missing, instead of reorganizing the overall data or declaring it invalid, web data extraction algorithms can help to enrich the databank from authorized and globally recognized web sources which can be seen as a huge change and an exciting way of research. Moreover, the biggest advantage of web data extraction is that the researchers and professional groups can conduct their own research studies without wasting their concern on data collection and its quality, searching for the particular keywords and saving the results manually. Particularly, the main motive of this paper is to extract reliable data and store them in a structured format.

For the extraction of data, some steps have been performed manually like preparing the specifications for the extraction process and this can be reused for applications of the same domain. This extraction process requires regular

monitoring of its validity because if the information website will be restructured it will no longer valid. This paper proved that data extraction and mapping are viable and the next job of this approach is to improve search and extraction capability. Data cleaning approach is essential for the practical uses and analytics of extracted data.

The data integration is based on the database schema mapping, for which the schema of web data is carefully analyzed and provides a common format for integration of extracted data. This makes the data integration easier.

Thus, it is possible to exploit and use the data automatically available on some web pages. However, some manual steps are necessary if one's intention is to extract reliable data to be exploited by other programs or software agents. The proposed method can also be very useful and helpful in conservation of biodiversity of our planet.

References

Adali S, Candan KS, Papakonstantinou Y, Subrahmanian VS (1996) Query caching and optimization in distributed mediator systems. ACM SIGMOD Rec 10(1145/235968):233327

Adelberg B (1998) NoDoSE—a tool for semi-automatically extracting structured and semistructured data from text documents. ACM SIGMOD Rec 10(1145/276305):276330

Apers PM (1995) Identifying internet-related database research. In East/west database workshop. Springer, London, pp 183–193

Arasu A, Garcia-Molina H, University S (2003) Extracting structured data from Web pages. In: Proceedings of the 2003 ACM SIGMOD international conference on on Management of data—SIGMOD'03. https://doi.org/10.1145/872797.872799

Arens Y, Knoblock Ca., Shen WM (1996) Query reformulation for dynamic information integration. J Intell Inform Syst. https://doi.org/10.1007/BF00122124

Arocena GO, Mendelzon AO (1999) WebOQL: Restructuring documents, databases, and webs. Theory and practice of object systems. https://doi.org/10.1002/(SICI)1096-9942(1999)5:3%3c127::AID-TAPO2%3e3.0.CO;2-X

Arora NK (2018a) Environmental sustainability—necessary for survival. Environ Sustain 1(1):1–2. https://doi.org/10.1007/s42398-018-0013-3

Arora NK (2018b) Biodiversity conservation for sustainable future. Environ Sustain 1(2):109–111. https://doi.org/10.1007/s42398-018-0023-1

Ballesteros-Mejia L, Kitching IJ, Jetz W, Nagel P, Beck J (2013) Mapping the biodiversity of tropical insects: Species richness and inventory completeness of African sphingid moths. Global Ecol Biogeogr. https://doi.org/10.1111/geb.12039

Batini C, Lenzerini M, Navathe SB (1986) A comparative analysis of methodologies for database schema integration. ACM Comput Surv 10(1145/27633):27634

Batista-Navarro R, Zerva C, Nguyen NTH, Ananiadou S (2017) A text mining-based framework for constructing an RDF-compliant biodiversity knowledge repository. In: Communications in computer and information science. https://doi.org/10.1007/978-3-319-55209-5_3

Baumgartner R, Baumgartner R, Flesca S, Gottlob G, Flesca S, Gottlob G (2001) Visual web information extraction with lixto. In: Proceedings of the international conference on very large data bases

Baumgartner R, Gatterbauer W, Gottlob G (2009) Web data extraction system. In: Encyclopedia of database systems (pp. 3465-3471). Springer, Boston

Beneventano D, Bergamaschi S, Guerra F, Vincini M (2003) Synthesizing an integrated ontology. In: IEEE Internet Computing. https://doi.org/10.1109/MIC.2003.1232517

Berners-Lee T (1998) Web Architecture from 50,000 feet. W3C. https://www.w3.org/DesignIssues/Architecture.html. Accessed 23 September 2018

Berners-Lee T, M F (1999) Weaving the web, the original design and ultimate destiny of the World Wide Web by its inventor. Harper Business San Francisco. https://doi.org/10.1109/TPC.2000.843652

Bernstein PA, Haas LM (2008) Information integration in the enterprise. Commun ACM 10(1145/1378727):1378745

Blagoderov V, Kitching IJ, Livermore L, Simonsen TJ, Smith VS (2012) No specimen left behind: Industrial scale digitization of natural history collections. ZooKeys. https://doi.org/10.3897/zookeys.209.3178

Blakeley JA (1997) Universal data access with OLE DB. In: Proceedings IEEE COMPCON 97. https://doi.org/10.1109/CMPCON.1997.584662

Bonczek RH, Holsapple CW, Whinston AB (1978) Aiding decision makers with a generalized data base management system: an application to inventory management. Decision Sci. https://doi.org/10.1111/j.1540-5915.1978.tb01381.x

Brin S, Motwani R, Page L, Winograd T (1998) What can you do with a web in your pocket? IEEE Data Eng Bull 21(2):37–47

Califf ME, Mooney RJ (1999) Relational learning of pattern-match rules for information extraction. Comput Linguist. https://doi.org/10.1162/153244304322972685

Calvanese D, De Giacomo G, Lenzerini M (2001) A Framework for Ontology Integration. In: Proc. of the 2001 Int. Semantic Web Working Symposium (SWWS 2001)

Ceccarelli T (1997) Towards a planning support system for communal areas in the Zambezi Valley, Zimbabwe: a multi criteria evaluation linking farm household analysis, land evaluation and geographic information systems

Chang CH, Kuo SC (2004) OLERA: Semisupervised Web-data extraction with visual support. IEEE Intell Syst. https://doi.org/10.1109/MIS.2004.71

Chang C, Lui SC (2001) IEPAD: Information extraction based on pattern discovery. In:Proceedings of the 10th international conference on World Wide Web—WWW. https://doi.org/10.1145/371920.372182

Chawathe S, Garcia-Molina H, Hammer J, Ireland K, Papakonstantinou Y, Ullman J, Widom J (1994) The TSIMMIS project: integration of heterogenous information sources. In:Proceedings of IPSJ conference

Chen PPS (1976) The entity-relationship model—toward a unified view of data. ACM Trans Datab Syst. https://doi.org/10.1145/320434.320440

Chidlovskii B, Ragetli J, de Rijke M (2000) Automatic wrapper generation for web search engines. In: International conference on web-age information management (pp. 399-410). Springer, Berlin

Choi N, Song I-Y, Han H (2006) A survey on ontology mapping. ACM SIGMOD Record 10(1145/1168092):1168097

Cohen WW (1998) Integration of heterogeneous databases without common domains using queries based on textual similarity. In: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. https://doi.org/10.1145/276305.276323

Crescenzi V, Mecca G (1998) Grammars have exceptions. Inform Syst. https://doi.org/10.1016/S0306-4379 (98)00028-3

Crescenzi V, Mecca G, Merialdo P (2001) Roadrunner: towards automatic data extraction from large web sites. In: Proceedings of the 27th International Conference on Very Large Data Bases

Date CJ (1995) An introduction to database systems. In: An introduction to database systems. https://doi.org/10.3145/epi.2009.jul.14

Doan A, Domingos P, Halevy A (2003) Learning to match the schemas of data sources: a multistrategy approach. Mach Learn. https://doi.org/10.1023/A:1021765902788

Drew P, King R, McLeod D, Rusinkiewicz M, Silberschatz A (1993) Report of the workshop on semantic heterogeneity and interpolation in multidatabase systems. {ACM} {SIGMOD} Record. https://doi.org/10.1145/163090.163098

Elmagarmid AK (1992) Database transaction models for advanced applications. Database

Embley DW, Campbell DM, Jiang YS, Liddle SW, Lonsdale DW, Ng YK, Smith RD (1999) Conceptual-model-based data extraction from multiple-record Web pages. Data Knowl Eng. https://doi.org/10.1016/S0169-023X(99)00027-0

Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) Knowledge discovery and data mining: towards a unifying framework. In: Int Conf on Knowledge Discovery and Data Mining

Fensel D, Van Harmelen F, Klein M, Akkermans H, Broekstra J, Fluit C, Krohn U (2000) On-to-knowledge: ontology-based tools for knowledge management. J Bus Ethic

Ferrara E, De Meo P, Fiumara G, Baumgartner R (2014) Web data extraction, applications and techniques: A survey. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2014.07.007

Finkelstein C (1989) An introduction to information engineering: from strategic planning to information systems. Addison-Wesley, Sydney, p 52

Florescu D, Levy AY, Mendelzon AO (1998) Database techniques for the World-Wide Web: a survey. SIGMOD Rec 27(3):59–74

Freitag D (2000) Machine learning for information extraction in informal domains. Mach Learn. https://doi.org/10.1023/A:1007601113994

Fridman N, Musen M (2000) PROMPT: Algorithm and tool for automated ontology merging and alignment. Proc. AAAI'00

Friedman M, Weld DS (1997) Efficiently executing information-gathering plans. In: In Proc. of the Int. Joint Conf. of AI (IJCAI

Fuxman A, Hernandez MA, Ho H, Miller RJ, Papotti P, Roma Tre U, Popa L (2006) Nested mappings: schema mapping reloaded. VLDB

Gangemi A, Guarino N, Masolo C, Oltramari A (2003) Sweetening WORDNET with DOLCE. AI magazine. https://doi.org/10.1007/3-540-45810-7

Gennari JH, Musen MA, Fergerson RW, Grosso WE, Crubezy M, Eriksson H, Tu SW (2003) The evolution of Protégé: An environment for knowledge-based systems development. International J Hum Comput Stud. https://doi.org/10.1016/S1071-5819(02)00127-1

Georgakopoulos D, Rusinkiewicz M, Sheth AP (1994) Using tickets to enforce the serializability of multidatabase Transactions. In: IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/69.273035

Ghawi R, Cullot N (2007) Database-to-ontology mapping generation for semantic interoperability. VDBL'07 Conference, VLDB Endowment ACM

Haas LM, Kossmann D, Wimmers EL, Yang J (1997) Optimizing queries across diverse data sources. Vldb

Halevy AY (2001) Answering queries using views: a survey. VLDB J. https://doi.org/10.1007/s007780100054

Halevy A, Ordille J (2006) Data integration : the teenage years. artificial intelligence. integration : the teenage yea

Hammer M, McLeod D (1979) On Database Management System Architecture (No. MIT/LCS/TM-141). MASSACHUSETTS INST OF TECH CAMBRIDGE LAB FOR COMPUTER SCIENCE

Hammer J, McHugh J, Garcia-Molina H (1997) Semistructured data: the TSIMMIS experience. In: Proceedings of the 1st east-european symposium on advances in databases and information systems (ADBIS)

Hardisty AR, Bacall F, Beard N, Balcázar-Vargas MP, Balech B, Barcza Z, Yilmaz P (2016) BioVeL: a virtual laboratory for data analysis and modelling in biodiversity science and ecology. BMC Ecol. https://doi.org/10.1186/s12898-016-0103-y

Heimbigner D, McLeod D (1985) A federated architecture for information management. ACM Trans Inform Syst 10(1145/4229):4233

Hogue A, Karger D (2005) Thresher : automating the unwrapping of semantic content from the World Wide Web. WWW'05: In: Proceedings of the 14th international conference on World Wide Web. https://doi.org/10.1145/1060745.1060762

Hsu CN, Dung MT (1998) Generating finite-state transducers for semi-structured data extraction from the Web. Inform Syst. https://doi.org/10.1016/S0306-4379

Huber GP (1990) A theory of the effects of advanced information technologies on organizational design, intelligence, and decision making. Acad Manag Rev. https://doi.org/10.2307/258105

Hull R (1997) Managing semantic heterogeneity in databases: a theoretical prospective. In: Proceedings of the sixteenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems (pp. 51-61). ACM

International Business Machines Corporation (1978) Business systems planning: Information systems planning guide. IBM

Irmak U, Suel T (2006) Interactive wrapper generation with minimal user effort. In: Proceedings of the 15th international conference on World Wide Web—WWW'06. https://doi.org/10.1145/1135777.1135859

IUCN Red list (2018) Numbers of threatened species by major groups of organisms (1996–2018). http://cmsdocs.s3.amazonaws.com/summarystats/2018-1_Summary_Stats_Page_Documents/2018_1_RL_Stats_Table_1.pdf. Accessed 20 September 2018

Jones AC (2006) Applying computer science research to biodiversity informatics: Some experiences and lessons. Trans Comput Syst Biol. https://doi.org/10.1007/11732488_4

Jones A, Xu X, Pittas N, Gray W, Fiddian N, White RJ, Brandt S (2000) SPICE: a flexible architecture for integrating autonomous databases to comprise a distributed catalogue of life. In: Database and expert systems applications, lecture notes in computer science. https://doi.org/10.1007/3-540-44469-6_92

Kadam VB, Pakle GK (2014) A survey on HTML structure aware and tree based web data scraping technique. Int J Comput Sci Inform Technol 5(2):1655–1658

Kalfoglou Y, SchorlemmeR M (2003) Ontology mapping: the state of the art. Knowl Eng Rev. https://doi.org/10.1017/S0269888903000651

Kayed M, Chang CH (2010) FiVaTech: Page-level web data extraction from template pages. In: IEEE Transactions on Knowledge and Data Engineering. https://doi.org/10.1109/TKDE.2009.82

Kossmann D (2000) The state of the art in distributed query processing. ACM Comput Surv 10(1145/371578):371598

Kushmerick N (2000) Wrapper induction: efficiency and expressiveness. Artif Intell. https://doi.org/10.1016/S0004-3702(99)00100-9

Laender AHF, Ribeiro-Neto BA, da Silva AS, Teixeira JS (2002a) A brief survey of web data extraction tools. ACM SIGMOD Rec 10(1145/565117):565137

Laender AHF, Ribeiro-Neto B, De Silva AS (2002) DEByE—Data extraction by example. Data Knowl Eng. https://doi.org/10.1016/S0169-023X(01)00047-7

Lage JP, Da Silva AS, Golgher PB, Laender AHF (2004) Automatic generation of agents for collecting hidden Web pages for data

extraction. In : Data and Knowledge Engineering. https://doi.org/10.1016/j.datak.2003.10.003
Levy Y, Rajaraman A, Ordille J (1996) Querying heterogeneous information sources using source descriptions. In: Proceeding VLDB'96 proceedings of the 22th international conference on very large data bases. https://doi.org/10.1049/tpe.1981.0030
Litwin W, Abdellatif A (1986) Multidatabase interoperability. Computer. https://doi.org/10.1109/MC.1986.1663123
Litwin W, Mark L, Roussopoulos N (1990) Interoperability of multiple autonomous databases. ACM Comput Surv 10(1145/96602):96608
Liu L, Pu C, Han W (2000) XWRAP : An XML—enabled wrapper construction system for web information sources. In: Proceedings of the 16th international conference on data engineering. https://doi.org/10.1109/ICDE.2000.839475
Malone TW, Yates J, Benjamin RI (1987) Electronic markets and electronic hierarchies. Commun ACM 10(1145/214762):214766
Martin J, Finkelstein C (1989) Information engineering. Prentice Hall, Englewood Cliffs
Mathew C, Güntsch A, Obst M, Vicario S, Haines R, Williams A, Goble C (2014) A semi-automated workflow for biodiversity data retrieval, cleaning, and quality control. Biodiv Data J. https://doi.org/10.3897/BDJ.2.e4221
McCarthy WE (1982) The REA accounting model—a generalized framework for accounting systems in a shared data environment. The Account Rev
Meersman R (2005) The use of lexicons and other computer-linguistic tools in semantics, design and cooperation of database systems. Star (2005)
Miller RJ, Haas LM, Hernández Ma (2000) Schema mapping as query discovery. In: Proceedings of the 26th international conference on very large data bases
Murphy J, Hashim NH, O'Connor P (2007) Take me back: validating the wayback machine. J Comput Med Commun. https://doi.org/10.1111/j.1083-6101.2007.00386.x
Muslea I, Minton S, Knoblock CA (2001) Hierarchical wrapper induction for semistructured information sources. Autonom Agent Multi-Agent Syst. https://doi.org/10.1023/A:1010022931168
Niles I, Pease A (2001) Towards a standard upper ontology. In: Proceedings of the international conference on formal ontology in information systems—FOIS'01. https://doi.org/10.1145/505168.505170
Nowak J, Nogueras-Iso J, Peedell S (2005) Issues of multilinguality in creating a European SDI-The perspective for spatial data interoperability. In: 11th ECGI GIS workshop ESDI setting the framework Alghero Sardinia
O'Sullivan B, Keady S, Keane E, Irwin S, O'Halloran J (2010). Data mining for biodiversity prediction in forests. In: Frontiers in artificial intelligence and applications. https://doi.org/10.3233/978-1-60750-606-5-289
Ouksel AM, Sheth A (1999) Semantic interoperability in global information systems. ACM Sigmod Record 28(1):5–12
Page RDM (2011) Extracting scientific articles from a large digital archive: BioStor and the biodiversity heritage library. BMC Bioinform. https://doi.org/10.1186/1471-2105-12-187
Raghavan S, Garcia-Molina H (2001) Integrating diverse information management systems: a brief survey. Technical Report, Stanford
Reis DC, Golgher PB, Silva AS, Laender AF (2004) Automatic web news extraction using tree edit distance. In: Proceedings of the 13th conference on World Wide Web - WWW'04. https://doi.org/10.1145/988672.988740
Ribeiro-Neto B, Laender aHF, Da Silva aS (1999) Extracting semi-structured data through examples. In: Proceedings of the Eighth International Conference on Information and Knowledge Management. https://doi.org/10.1145/319950.319962
Roy PS, Karnatak H, Kushwaha SPS, Roy A, Saran S (2012) India's plant diversity database at landscape level on geospatial platform: prospects and utility in today's changing climate. Curr Sci 102(8):1136–1142
Sahuguet A, Azavant F (2001) Building intelligent Web applications using lightweight wrappers. Data Knowl Eng. https://doi.org/10.1016/S0169-023X(00)00051-3
Saran S, Kushwaha SPS, Ganeshaiah KN, Roy PS, Murthy YK (2012) Indian Bioresource Information Network (IBIN): a distributed bioresource national portal. ISG Newslett 18(3):6
Selkow SM (1977) The tree-to-tree editing problem. Inform Proces Lett. https://doi.org/10.1016/0020-0190(77)90064-3
Shanmughavel P (2007) An overview on biodiversity information in databases. Bioinformation. https://doi.org/10.6026/97320630001367
Shekhar S (2004) Spatial data mining and geo-spatial interoperability. In Report of the NCGIA specialist meeting on spatial webs, Santa Barbara, December 2–4 2004
Sheth AP (1999) Changing focus on interoperability in information systems:from system, syntax, structure to semantics. In: Interoperating geographic information systems. https://doi.org/10.1007/978-1-4615-5189-8_2
Sheth A, Kashyap V (1993) So far (schematically) yet so near (semantically). In interoperable database. Systems 5:283–312
Sheth AP, Larson JA (1990) Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Comput Surv 10(1145/96602):96604
Silva N, Rocha J (2003) Ontology mapping for interoperability in semantic web. In: ICWI, pp. 603–610
Silvertown J (2009) A new dawn for citizen science. Trends Ecol Evol. https://doi.org/10.1016/j.tree.2009.03.017
Singh P, Saran S, Kumar D, Padalia H, Srivastava A, Kumar AS (2018) Species mapping using citizen science approach through IBIN portal: use case in foothills of Himalaya. J Indian Soc Remote Sens, 1–13
Smedt TD, Daelemans W (2012) Pattern for python. J Mach Learn Res 13:2063–2067
Soderland S (1999) Learning information extraction rules for semi-structured and free text. Mach Learn. https://doi.org/10.1023/A:1007562322031
Sonsilphong S, Arch-int N, Arch-int S (2012) Rule-based semantic web services annotation for healthcare information integration. In: Computing and networking technology (iccnt), 2012 8th international conference on (pp. 147-152). IEEE
Species 2000 Secretariat (2009) Species 2000. http://www.sp2000.org/index.php?option=com_content&task=view&id=40&Itemid=49. Accessed 23 September 2018
Stonebraker M, Aoki PM, Litwin W, Pfeffer A, Sah A, Sidell J, Yu A (1996) Mariposa: a wide-area distributed database system. VLDB J. https://doi.org/10.1007/s007780050015
Tomasic A, Raschid L, Valduriez P (1998) Scaling access to heterogeneous data sources with DISCO. In: IEEE transactions on knowledge and data engineering. https://doi.org/10.1109/69.729736
Ullman JD (2000) Information integration using logical views. Theor Comput Sci. https://doi.org/10.1016/S0304-3975(99)00219-4
Veltman KH (2001) Syntactic and semantic interoperability: new approaches to knowledge and the semantic web. N Rev Inform Netw 7(1):159–183
Wache H, Vögele T, Visser U, Stuckenschmidt H, Schuster G, Neumann H, Hübner S (2001) Ontology-based integration of information-a survey of existing approaches. IJCAI Workshop: Ontologies and Information Sharing

Wang J, Lochovsky FH (2003) Data extraction and label assignment for web databases. In: Proceedings of the twelfth international conference on World Wide Web—WWW'03. https://doi.org/10.1145/775152.775179

Wilson PS (2007) What mapping and modeling means to the HIM professional. Perspectives in Health Information Management/AHIMA, American Health Information Management Association, Chicago, p 4

Woelk D, Bohrer B, Jacobs N, Ong K, Tomlinson C, Unnikrishnan C (1995) Carnot and InfoSleuth: database technology and the world wide web. In: ACM SIGMOD Record (Vol. 24, No. 2, pp. 443-444). ACM

Yang W (1991) Identifying syntactic differences between two programs. software: practice and experience. https://doi.org/10.1002/spe.4380210706

Yesson C, Brewer PW, Sutton T, Caithness N, Pahwa JS, Burgess M, Culham A (2007) How global is the global biodiversity information facility? PLoS One. https://doi.org/10.1371/journal.pone.0001124

Zhai Y, Liu B (2005) Web data extraction based on partial tree alignment. In: Proceedings of the 14th international conference on World Wide Web—WWW'05. https://doi.org/10.1145/1060745.1060761

Zhai Y, Liu B (2006) Structured data extraction from the web based on partial tree alignment. In: IEEE transactions on knowledge and data engineering. https://doi.org/10.1109/TKDE.2006.197

Zhao H, Zhang S, Zhou J, Wang M. (2007). Semantic model based heterogeneous databases integration platform. In: Icnc (pp. 366-370). IEEE