Environmental Challenges 3 (2021) 100064

Contents lists available at ScienceDirect

Environmental Challenges

journal homepage: www.elsevier.com/locate/envc

A semantic approach to enable data integration for the domain of flood risk management

Vatsala Nundloll[a,*], Rob Lamb[a,b], Barry Hankin[c], Gordon Blair[a]

[a] Lancaster University, Lancaster, UK
[b] JBA Trust, Skipton, UK
[c] JBA Consulting Ltd., Skipton, UK

ARTICLE INFO

Keywords:
Ontologies
Structured data
Unstructured data
Semantic integration
Natural language processing
Flood risk management

ABSTRACT

With so many things around us continuously producing and processing data, be it mobile phones, or sensors attached to devices, or satellites sitting thousands of kilometres above our heads, data is becoming increasingly heterogeneous. Scientists are inevitably faced with data challenges, coined as the 4 V's of data - volume, variety, velocity and veracity. In this paper, we address the issue of data variety. The task of integrating and querying such heterogeneous data is further compounded if the data is in unstructured form. We hence propose an approach using Semantic Web and Natural Language Processing techniques to resolve the heterogeneity arising in data formats, bring together structured and unstructured data and provide a unified data model to query from disparate data sets.

1. Introduction

Recent advances in technology have led to an explosion in the availability of data, often referred to as 'big data'. This is particularly true in the domain of flood risk management, where there is a plethora of data now available from a range of sources including from satellite imagery, ground-based sensors, citizen science and data mined from the web. To be effective, however, this data needs to be brought together to achieve the necessary level of integration and subsequently to foster decision-making that is more open and transparent and that supports the necessary level of collaboration amongst different stakeholders and specialists in this important area. In contrast, at present, this data is often siloed and this is increasingly a major problem for the field. This siloing problem is easily overcome by bringing the data together in one place, for example exploiting the potential of cloud computing. This is a necessary first step but this alone is not sufficient - there is also a need to tackle a key problem of data heterogeneity. Heterogeneity in data, either in the form of diverse data formats or in the form of disparate data sources, poses a serious hindrance to domain specialists in moving towards a more integrated and collaborative mode of working. As a result, there is an urgent need to tackle this data heterogeneity problem in flood risk management (note that this is also one of the 4 V's of 'big data', i.e. 'variety', which sits alongside - 'volume', 'velocity' and 'veracity'), and to build a unified model that can facilitate the integration of disparate data sets.

The task of integrating data is further compounded with unstructured forms of data. Many existing approaches talk about integrating structured data mostly. With technological advances in Natural Language Processing (NLP) methods, there is ongoing research on extracting information from unstructured data such as text and images. However, there is little attention on how to bring unstructured data together with structured data, which is a key focus of this paper. Existing literature on the use of NLP to extract information from unstructured text focuses on information that is textual in nature such as a person's name, nationality, etc. However, in the flood domain, the information to be extracted also consists of important quantitative information such as measurements of a river level, flow of water, etc. This information is more challenging as we need to preserve the information context. Hence, we not only use NLP techniques to extract information from text, but also reflect on the use of an ontological approach to preserve the context of the information extracted and to integrate information from structured and unstructured sources in a more meaningful way. Hence, this paper presents an approach using Natural Language Processing alongside Semantic Web technologies to extract unstructured information from text and to bring this together with information from more structured sources with the aim to make more informed queries.

Bringing these threads together, this paper investigates possible technological solutions to achieve data integration for the field of flood risk management. This breaks down into the following research questions:

[*] Corresponding author.
E-mail addresses: vatsala@lancaster.ac.uk (V. Nundloll), rob.lamb@jbatrust.org (R. Lamb), barry.hankin@jbaconsulting.com (B. Hankin), g.blair@lancaster.ac.uk (G. Blair).

https://doi.org/10.1016/j.envc.2021.100064
Received 26 November 2020; Received in revised form 22 February 2021; Accepted 23 February 2021

1. Can we achieve full and meaningful data integration in flood risk management through the application of Semantic Web technologies, including the use of domain ontologies, and does this achieve our goal of a unified view over all available flood risk management data;
2. Can we naturally extend the above solution to unify both structured and unstructured data sources, through an approach based on Natural Language Processing;
3. Can we subsequently interrogate and query this data as a unified whole, drawing data from disparate sources including structured and unstructured sources, and hence support our vision of enriched support for collaborative and transparent decision-making in flood risk management.

The work presented has been carried out in collaboration with domain experts from the field of flood risk management and is part of a transdisciplinary project investigating the use of digital technologies to mitigate the challenges faced in the environmental domain.

The paper is structured as follows. Section 2 presents related work, considering literature on data integration particularly in the domain of flood risk management. Section 3 then gives a brief introduction to ontologies. Section 4 describes the methodology used and the overall approach adopted, and Section 5 presents the prototype design. Section 6 provides an evaluation of the proposed approach, and Section 7 presents a discussion emanating from this analysis; Finally Section 8 concludes the paper, including statements on future research directions in this promising field.

2. Related work

Data integration is the process of bringing together data from disparate sources to enable a unified query mechanism. Authors in Towe et al. (2020) highlight the need to bring data together from diverse sources for drawing better analysis in flood risk management and bring to attention the inherent dependency of decision-making in flood risk management on data. With the added challenge that data is provided from heterogeneous sources, the key requirement is to enable the integration and subsequent analyses of a complex array of data in order to tackle the challenges in flood modelling. The authors present the concept of a hypercube model — using cloud computing, data integration using a semantic web approach and the use of notebook technologies to help bring more insight into flood modelling. Our paper provides a detailed narrative of the semantic data integration work presented in Towe et al. (2020) and presents a linked data model using Semantic Web and Natural Language Processing techniques to address the data heterogeneity issue arising in the flood domain.

Moreover, the authors in Blair et al. (2019) envision the need to reduce uncertainty in environmental models through generic models capable of capturing processes and behaviours across multiple places. The core motivation of such models is to exploit as much knowledge as possible regarding a particular place in order to reduce this uncertainty. Blair et al. (2019) also highlights the need to bring data together from heterogeneous sources in order to draw better insight from process and data models. For example, the knowledge for drawing flood predictions can be provided from remote sensing data, historical Parish records, flood marks, satellite imagery and local sensors. New data mining techniques enable scraping the web or social media for information in the form of text and images. Better insight can be drawn from these data sources if combined together for usage. However, the heterogeneity arising in the data formats and data sources poses a barrier to scientists to properly utilise data from disparate sources, thus causing data to remain silo-ed. Furthermore, Beven and Smith (2015) emphasises the need to validate information in hydrological modelling, emanating from the fact that modelling in hydrology is limited by its measurement techniques. The inconsistencies arising in data introduce disinformation into the models. The main sources of uncertainty in these models are due to errors in the models, errors in the observations in model calibration and computational constraints. The problem highlighted in this paper calls for the need to bring in other dimensions of data, other than just the observed measurements, that can possibly help to alleviate the inconsistencies arising in the model outputs. The Environment Agency in England, for instance, owning hydrology models on flood defences, flood risk assessments, detailed constructions of drainage systems, etc., recognises the necessity for these models to evolve over time by incorporating new data and bringing in technical improvements to their models. The agency published a technical report (Environment-Agency-Cost, 2018) drawing estimates of economic cost of flood events to highlight the devastating impact on property damage that can be caused as a result of significant flood events. There is also a growing recognition of the importance of drawing new types of data in future flood risk assessments (Environment-Agency, 2020). The agency stresses there remains a significant gap in literature and practice as to how new data sources can be brought in together in order to derive better process-driven and data-driven paradigms. Baldassarre et al. (2016) and Smith et al. (2017) advocate the need to bring in heterogeneous data to improve the hydrology models and to mitigate risks. Smith et al. (2017), for instance, talks about a real-time modelling framework using information gathered from social media to identify flood-prone areas. McCallum et al. (2016) advocates the use of new digital technologies to collect and analyse data, provided by communities and citizens across the globe, that may be used to monitor, validate and reduce flood risks.

The data integration problem points to the need to tackle the data heterogeneity (variety) issue first. The Semantic Web (SemanticWeb, 2021) is a vision of the World Wide Web Consortium to build a web of data which are linked from different sources. Semantic technologies such as RDF, OWL, IRI and SPARQL have been proposed to enable data integration. Another valuable contribution from the Semantic Web is the use of ontologies which help to resolve the data variety problem by enforcing data standardisation and hence enabling the interoperability of heterogeneous data. A good introduction of Semantic Data integration is provided by ontotext (a leading company on the use of Semantic Web techniques) (Ontotext, 2021a). A survey of a few existing ontologies can be found at d'Aquin and Noy (2012) and methodological guidelines for reusing ontologies have been proposed in Fernández-López et al. (2013). Ontologies have been widely applied in bioinformatics and healthcare information systems. Ison et al. (2013) shows the need to integrate information about bioinformatics operations through the use of ontologies. He et al. (2017) presents innovative semantic-based methods to address important problems in healthcare. Zhang et al. (2018) introduces an ontological framework to support integration of cancer-related data. Turki et al. (2019) is a large-scale medical database using ontologies to resolve heterogeneity issues and to bring data together from different biomedical systems. Other domains where ontologies are being applied are Web Search, Ecommerce, Geographic Information Systems, etc. For instance, Sun et al. (2019) presents the use of ontologies to tackle semantic heterogeneity and integration of geospatial data. Best practices for ontology engineering in the IoT domain can be found at Atemezing (2015).

The need for data integration for the flood risk management domain shows the scope for using ontologies in this field to bring in disparate data sources together. However, the complexity of integrating data is further compounded with some data being in unstructured form. With the rise in social media platforms, there is a lot of citizen science information available in form of text and images. A lot of scientific findings are also available in textual documents. Past literature such as Montori, F. et al. (2016), Ziegler and Dittrich (2004), Araque et al. (2008) and Bernadette et al. (2002) have proposed data integration solutions restricted to structured data only but the authors recognise the value of folding in unstructured data as well. For instance, Demir and Krajewski (2013) demonstrates a flood warning system that builds on sensor data collected from different sources and integrated so that they can be shared in common data formats. However, this integrated platform considers structured data only. The reason why unstructured data has lagged behind has been due to major technological barriers.

However, we can see a few commercial online products such as (IBM, 2020, Astera, 2020) and Ontotext (Ontotext, 2021b) that provide services for extracting information from unstructured data. They may be efficient tools but can be costly and their price may vary depending on the size of data, complexity of data etc. Moreover, the client needs to pay each time they want to extract information from a new document. However, with recent technological advances in open source technologies such as Natural Language Processing frameworks, often exploiting new developments in Machine Learning, the ability to extract information from unstructured sources is increasing. Machine Learning is the science of training computers to learn and to draw predictions whilst Natural Language Processing is the science of understanding the interactions between the computer and the human language. There is ongoing research in this area and one prominent area is in the use of Sentiment Analysis to classify the opinions of people from customer surveys, reviews and social media posts. The work presented in Horecki and Mazurkiewicz (2015) talks about the use of Natural Language Processing to classify words derived from text into emotions or opinions. Text classification is also rapidly gaining popularity. Although Romanov et al. (2019), for instance, has applied text pre-processing, feature extraction and classification of the extracted features using different Machine Learning models, the authors recognise the need for more research contributions in drawing information from textual sources.

3. Background on ontologies

What is an ontology? Ontologies (Gruber, 1993) are formal explicit specifications of a shared conceptualisation. They model some aspect of the world (called a domain), and provide a simplified view of some phenomenon in the world that we want to represent. A domain can be defined as any aspect related to the world, for example an educational domain, a medical domain etc. The ontology acts as a vocabulary to explicitly define concepts, properties, relations, functions, constraints of a particular domain and also represents the schema of the data being modelled. It can also enable to uniquely identify each concept through an IRI (Internationalized Resource Identifier). For example, for a bird domain, an ontology can be used to define the features of a bird such as the feathers, wings, beak, eggs, etc. and assign an IRI to each bird concept. A popular open-source software available for the design of ontologies is Protege (2020). We can add a layer of metadata above ‘raw’ data by enriching each data atom with an ontological concept having a definition. This metadata layer can thus enable to abstract over disparate data sources and enable their integration. For example, if bird feathers are labelled as *Feather* and *BirdFeather* in two separate datasets, they can still be annotated with the same ontological concept about bird feathers. This abstraction can enable to bring both datasets together.

Instead of devoting a lot of time and effort in designing one global ontology to represent the domain concepts and its associated schema, it is generally recommended to maintain a separation of concerns between the data and the domain concepts. This approach can help a better manipulation and maintenance of the ontologies used. Hence, a hybrid approach can be adopted, where the general concepts of the flood domain are represented through a **flood domain ontology**; and the schema of the flood-related data are represented through a single or multiple **data ontologies**. Since it is not tied to any data, the flood risk management domain ontology can be re-used with other flood-related data sets.

Brief context on domain and data ontologies Different ontology configurations can be used to enable data integration - single ontology, multiple ontology and hybrid ontology approaches. A **single ontology** approach requires all the source schemas to be directly related to a shared global ontology. Examples of such systems are SIMS (Arens et al., 1996), Carnot (Collet et al., 1991) and PICSEL (Goasdoué and Reynaud, 1999). The drawback of this approach is that the domain information needs to be updated every time there is a new information source. A **multiple ontology** approach requires every data source to be described by its own local ontology, and all the local ontologies are eventually mapped to each other. An example of such a system is the OBSERVER system (Mena et al., 2000). A **hybrid ontology** approach is a combination of the first two approaches - a domain ontology captures the domain knowledge at a level of abstraction free from implementation concerns; a data ontology models the structure of a particular dataset and is used to interface between data and a domain ontology. This hybrid architecture provides greater flexibility in integrating new data sources as they can be represented using local ontologies. This view is also backed by Cruz and Xiao (2003), Dolbear et al. (2005).

4. Methodology

We adopt an agile approach as the core methodology to underpin this research. An agile approach is one which is done in an iterative way, where each step can be revisited and altered as per the needs of a process. It encapsulates a range of principles and values expressed in software development. This approach helps to alleviate the inefficient traditional practices in software development where continuous meetings, heavy documentation, strict adherence to a gantt chart used to be the norm. The result is a set of agile methods iterating through a continuous delivery of software in order to reach a solution. A good introduction to this approach is provided at Ferrario et al. (2014). This agile process started off with a workshop where the inter-disciplinary project partners, researchers and stakeholders from the flood community met to discuss and reflect about the challenges involved in the flood domain. The main output resulting from this workshop was the need for a more data driven approach to tackle the flood challenges. As a consequence, we decided to adopt the agile method as our research methodology in order to investigate into technologies that can help to bridge the gap between structured and unstructured data forms and bring heterogeneous datasets together into a unified model that can be eventually queried. The technologies that have been investigated using this methodology are: (i) Natural Language Processing (NLP) techniques to extract information from an unstructured data source; (ii) Semantic Web techniques to bridge the gap between structured and unstructured data forms and enable their integration. Ongoing research on the use of NLP techniques to extract information from textual sources is currently mostly around identifying textual entities from text. But in the flood context, crucial information in the form of numerical quantities is also an important feature of flood documents. We need to be able to extract these measurement data without losing their context. Hence, the methodology proposed looks at how to extract textual as well as numerical quantities in such a way that we can still preserve their context, and how to bring this information together with related information from disparate data sets. The suggested architecture to enable the data integration can be summed up as follows:

1. Data Provisioning for a given domain
2. Data Integration of structured and unstructured sources into a linked data model
3. Data Querying from the linked data model

4.1. Data Provisioning for a given domain

The first step involves procuring the different data sources that require integrating. Structured data are ‘ready’ to be used owing to their ’field label/value’ representation. They may need to undergo through some data processing or cleaning phase though. For the flood domain, structured data takes the form of model outputs or can be available as a rich array of measurement data ranging from instruments, satellite imagery through to remote sensing. This complexity is further compounded with unstructured forms of data such as text, images or even hand-written notes. For unstructured data, they need to go through an extra step where the required information needs to be extracted using Natural Language Processing techniques.

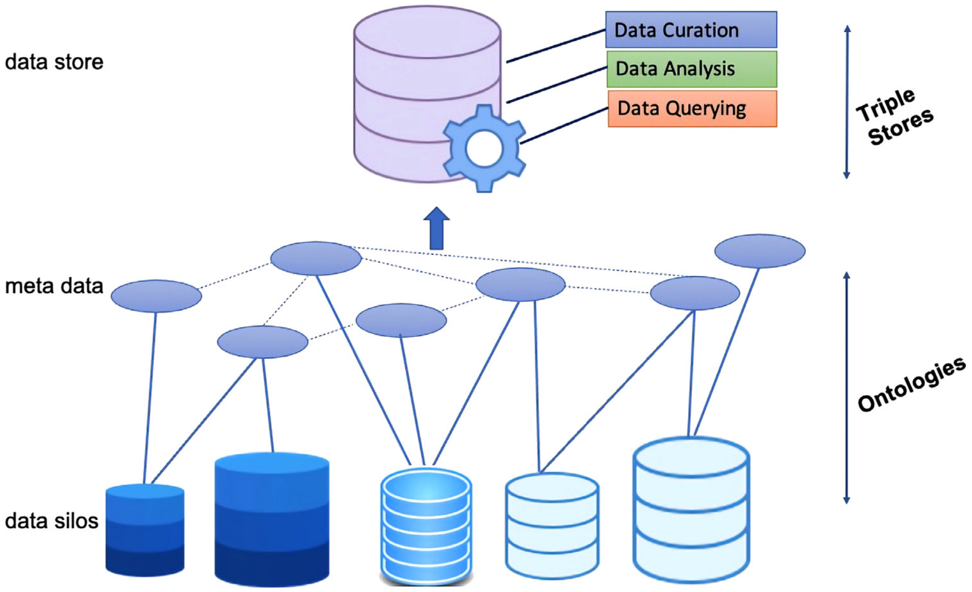


Fig. 1. Data Integration using ontologies.

4.2. *Data Integration of structured and unstructured sources into a linked data model*

This step is broken down into the ontology design for the given domain first followed by a semantic enrichment of the data using ontological concepts.

4.2.1. *Ontology design for the given domain*

Fig. 1 shows how ontologies create that extra metadata layer that sits on top of disparate data sources, and enable data to be integrated into a unified model. This metadata layer enables the information sitting in the data stores to be abstracted and represented in a standardized format, that can enable the different sources to be integrated as shown. The integrated model can then be stored into a semantic data store (also known as a triple store) that can be queried and that can produce richer query results provided from different repositories. A hybrid ontological approach has been adopted here where the flood domain has been modelled through a domain ontology and the flood datasets have been modelled through data ontologies (see Section 3).

4.2.2. *Semantic Enrichment of data*

This step involves rendering the heterogeneous data into a standardized format so that they can be linked together. This standardized format is known as RDF (Resource Description Framework (RDF, 2014), a data representation format that enables ontologies to define data concepts in a standard way which is a triple format: **Subject-Predicate-Object**. The *Subject* and *Object* are the data components of a domain whilst the *Predicate* is the relationship between them. The ontology enrichment of ‘raw’ data is known as semantically-enriched data, and allows different data components to be annotated with relevant concepts from the ontology. Referring to the bird example, using bird observation data - such as the colour of the bird (BirdColour=Red) or shape of the beak (BeakShape=Sharp) - the semantically enriched data can look as follows: *Bird hasColour Red; Beak hasShape Sharp*. Each component in the triple is a concept from an ontology. For example, *‘Bird*’ and *‘Beak’* are both data concepts whilst *‘hasColour’* and *‘hasShape’* are relationship concepts. The values ‘*Red*’ and ‘*Sharp*’ can either be further data concepts or simply values especially if they are numerical. This semantic enrichment helps to abstract over the heterogeneous datasets and to bring them together into a linked data model as shown in Fig. 1.

4.2.3. Information Extraction from unstructured sources using NLP techniques

In this paper, we are considering text only as the unstructured form of data. As mentioned above, data needs to be standardised into RDF form before the integration step. Whilst this can be a relatively straightforward process for structured data, the unstructured data will need to go through an additional processing step, which is extracting the information first using Natural Language Processing methods. Given that RDF is a *Subject-Predicate-Object*(SPO) triple structure, the information extracted from text should be close enough to this triple structure so that it can easily be converted into RDF. Therefore, the idea is to extract a ‘subject’, ‘predicate’ and an ‘object’ from every sentence occurring in the text. In this way, even if the object component of a sentence contains numerical quantities, the latter does not lose its context since it is associated to a subject and a predicate, both also components from the same sentence.

4.3. *Data Querying from the linked data model*

Once the data has been semantically converted into RDF, they can be loaded as triples onto a semantic data store or triple store. The triple store enables data from different sources to be linked together, and this is known as a semantic linked data model. Triple stores can usually contain millions of triples and are licensed either as a desktop version or a cloud version, depending on the size of the data to be integrated. They also provide a query facility (through a semantic query language) to query the information stored and bring forward information queried from disparate datasets.

5. Prototype design

The prototype has been designed to validate the methodology proposed and is based on bringing flood-related data together. The idea of enabling data integration for the flood domain emanates from the need to do better hydrological modelling, as mentioned in Blair et al. (2019) and Towe et al. (2020) (see Section 2) emphasizing the need to bring in data from different sources to fold in other parameters to address the uncertainty issue in modelling. This work has been carried out in close collaboration with flood scientists from the Environment Agency(EA) (Environment-Agency, 2021), JBA Trust (JBATrust, 2021) and JBA Consulting group (JBAConsulting, 2021). The flood scientists

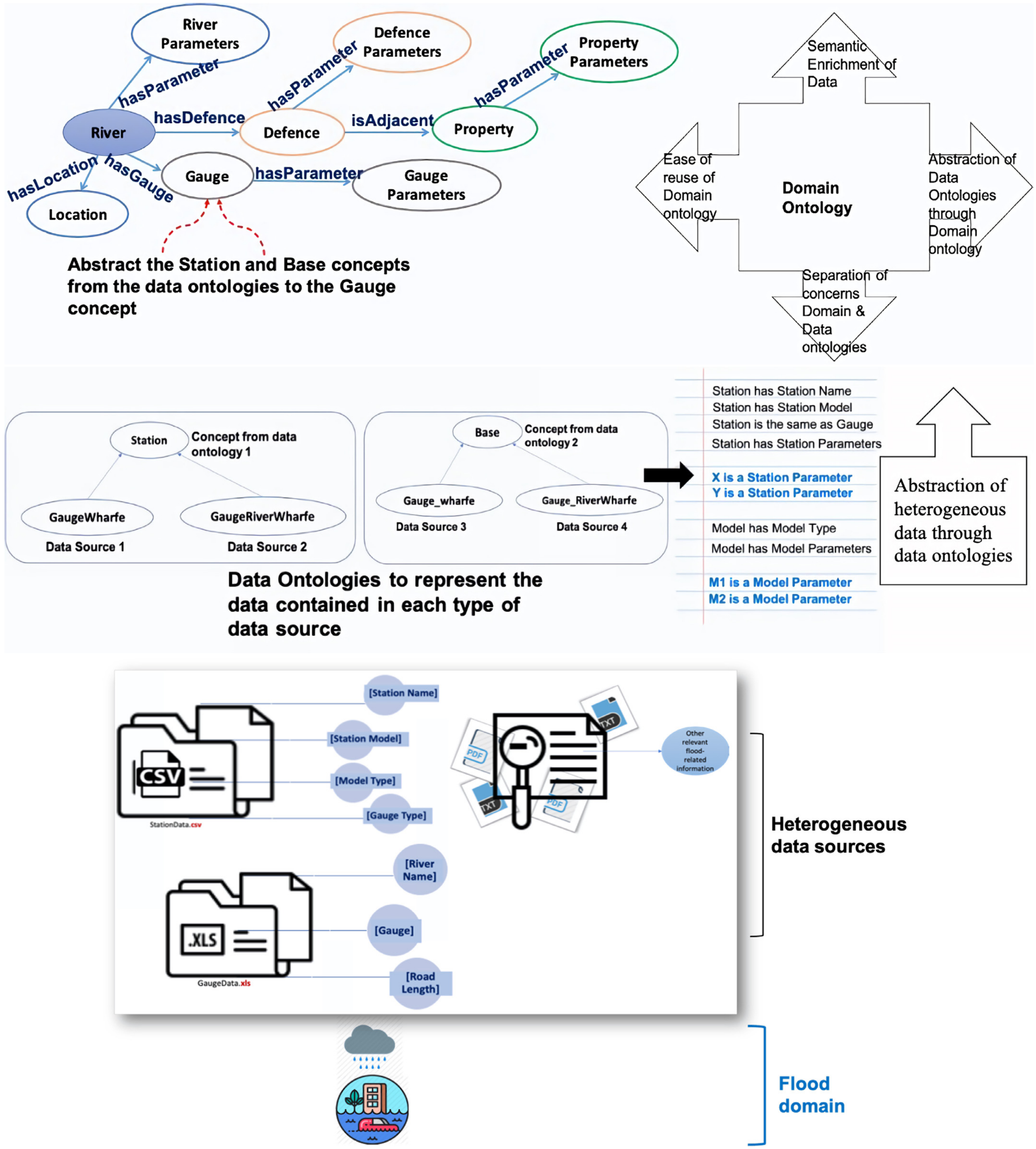

River Parameters
hasParameter
Defence Parameters
hasParameter
Property Parameters
hasParameter
River
hasDefence
Defence
isAdjacent
Property
hasLocation
hasGauge
Gauge
hasParameter
Gauge Parameters
Location
Abstract the Station and Base concepts from the data ontologies to the Gauge concept
Semantic Enrichment of Data
Ease of reuse of Domain ontology
Domain Ontology
Abstraction of Data Ontologies through Domain ontology
Separation of concerns Domain & Data ontologies
Station
Concept from data ontology 1
GaugeWharfe
GaugeRiverWharfe
Data Source 1
Data Source 2
Base
Concept from data ontology 2
Gauge_wharfe
Gauge_RiverWharfe
Data Source 3
Data Source 4
Data Ontologies to represent the data contained in each type of data source
Station has Station Name
Station has Station Model
Station is the same as Gauge
Station has Station Parameters
X is a Station Parameter
Y is a Station Parameter
Model has Model Type
Model has Model Parameters
M1 is a Model Parameter
M2 is a Model Parameter
Abstraction of heterogeneous data through data ontologies
[Station Name]
[Station Model]
[Model Type]
[Gauge Type]
CSV
StationData.csv
PDF
TXT
Other relevant flood-related information
[River Name]
[Gauge]
[Road Length]
.XLS
GaugeData.xls
Heterogeneous data sources
Flood domain

Fig. 2. Data integration using data and domain ontologies.

rely on measurement data about river levels, gauges, etc. to assess the impact of flooding on properties in flood-prone areas. However, they mention that there is a lot of useful information sitting in government reports that can help in their risk analysis and impact assessment. They expressed a need to draw information from one such government report and bring it together with the measurement data regarding the flood-prone areas in the UK. The challenge is how to extract this information and integrate it with the measurement data. Fig. 2 illustrates the prototype model, and the steps in the figure (bottom-up) are explained below.

5.1. Data source identification

5.1.1. Communities at risk data

The agencies EA and JBA Trust/Consulting provided the "Communities at Risk" dataset - a set of measurement data which was collected from flood-prone localities and was used to assess the impact of flooding in these areas. The data was provided in Excel files and was mainly about river gauges and building properties found in flood-prone areas. We considered two Excel files for this prototype. One Excel file contained information on the names of gauges at different river locations, the types of river gauges, gauge measurement data and details about hydraulic models used for analysis. The other Excel file contained details of properties (pricing, dimensions, types) found near those river gauges and details of the models used for analysis purposes. To complicate things, the details about these gauges and these residential/commercial properties were scattered across different spreadsheets in the files. This posed a hindrance to the flood experts in collating this information and using it in a timely manner for analysis purposes.

5.1.2. The Section19 report

There is a lot of information in government reports regarding flood-affected areas that are of relevance to the flood scientists in their day-to-day work. One such document is known as the Section19 report (FloodReport, 2021) and is circulated by the Risk Management Authorities in the UK in response to flooding events. These reports are normally in PDF, and represent a rich source of quantitative as well as qualitative information gathered by different stakeholders following flood events. They are considered to be an important source of flood risk management-related information for stakeholders involved in mitigating flood risks in flood-prone areas. For example, these reports list flood occurrences and their locations, the severity of the events, highlights the statutory responsibilities and duties of flood-risk management authorities, and denotes the types of emergency response schemes available to flood-prone areas amongst other relevant information. However, given the highly unstructured nature of this data source, it is difficult for the flood scientists to make adequate use of this information.

5.1.3. The importance of integrating the Section19 report and the Communities At Risk data

The provisioning of any vital information on flood-affected places can help the flood scientists to get a better understanding of these places. Bringing the "Communities At Risk" data and the Section19 report together can highlight all relevant details about flood-prone areas and can help give a better insight about these areas. Given the textual nature of the Section19 report, we need to find a way to extract the required information, combine it with the measurement data from the spreadsheets and query for information from the multi-source data model. This is a challenging process, and we believe that a combination of Semantic Web and Natural Language Processing techniques can help to bridge the gap between these heterogeneous sources and combine them into a unified model, known as a linked data model, which can then be queried.

5.2. Ontology design

As explained in the methodology section, the ontology is designed to stitch disparate data sets together by creating an abstraction layer above these data sets that binds different data components into relationships that reflect the different aspects of the flood domain. The result is the formation of a linked data model that links all the disparate flood-related data sets involved. Therefore, the first step in creating this linked data model for the flood domain is to have an ontology that reflects the different aspects of that domain. This domain ontology will model the different concepts and the relationships between the concepts. The knowledge about these concepts/relationships comes from the flood scientists who are the domain experts. Furthermore, the concepts are elicited by also scanning through the datasets to identify prominent entities, and through discussion with the scientists to determine their relevance. Further knowledge about the domain knowledge has also been gathered through literature survey. A hybrid ontological approach is adopted where a domain ontology captures the general concepts of the flood domain and data ontologies are used to model the different datasets that have been made available.

5.2.1. Domain ontology - EIA ontology

In order to support the practice of reusing existing ontologies, the EIA ontology (Environmental Impact Assessment) Garrido and Requena (2011) has been used to capture the flood domain concepts. The EIA ontology broadly shows the impact of human activities and natural phenomena on the environment; categorizes the impact identifiers; describes environmental factors, environmental services and impact assessment procedures. We found that the flood concept can been categorized as a natural phenomenon, which impacts on the environment and the lives of people. The reason for choosing the EIA ontology is the broad spectrum of environmental concepts it captures. This can allow to accommodate new sources of data from other parts of the ecosystem and this can help scientists to draw insight on the impact of flooding on a wider scale. Although a new ontology can be designed to represent the different aspects of the flood domain, we did not want to add yet another level of heterogeneity by contributing new ontologies when we can accommodate new concepts to an already existing one. Fig. 3 shows a snapshot of the EIA ontology, depicting a flood concept which occurs in a catchment area. There is a *'model'* concept to represent simulation/statistical/process models that can be utilized to analyse flood risks. This concept also has a set of input and output parameters, each defined as two separate concepts. The catchment area, on the other hand, is described as having:(i) a defence infrastructure - refers to the types of defence systems put in place to prevent flooding; and (ii) a receptor infrastructure - refers to the properties, structures, or landmarks found near a flood-prone area. The ontology is a reflection of the source-pathway-receptor approach commonly used in risk assessment and helps to uncover the different types of pathways leading to flooding and types of receptors impacting from this flooding. The ontology highlights what kinds of data need to be folded in so as to further contemplate the source-pathway-receptor approach. Our domain ontology is named *floodmodel.owl*, and shows the relationships between the different concepts identified for the flood domain, as depicted in Fig. 3.

5.2.2. Data ontology

The data ontology is simply a schematic representation of the flood-related data sets. We have created two data ontologies for modelling the *Communities at Risk* dataset and the data extracted from the *Section19* report. The Communities at Risk data ontology represents all possible parameters normally used in flood measurements such as gauge level, river velocity, river depth, etc. It has been named D***amage***A***ssessment.owl*** and is shown in Fig. 4 (a). Concepts such as *Gauge, Model, River, Property, etc.* in this figure represent the measurement parameters for the flood domain. The concepts can be further expanded to accommodate more fine-grained details which can be classified as sub-concepts. Regarding the data ontology for the Section19 report, it defines concepts such as 'Data Provenance' (data contributor, data collection methods, data source, data collected for, etc.), 'Observation Types' and 'Timestamp'. This ontology is called ***Localknowledge.owl***, shown in Fig. 4(b), and can be reused to represent any other data of a similar nature.

5.3. Semantic enrichment of structured data

5.3.1. Data transformation for structured data

When people talk about structured data, they are generally referring to data stored in RDBMS, Excel or CSV files. Such data are stored in a structured way, in a classic field/value pair. Nonetheless, there is still

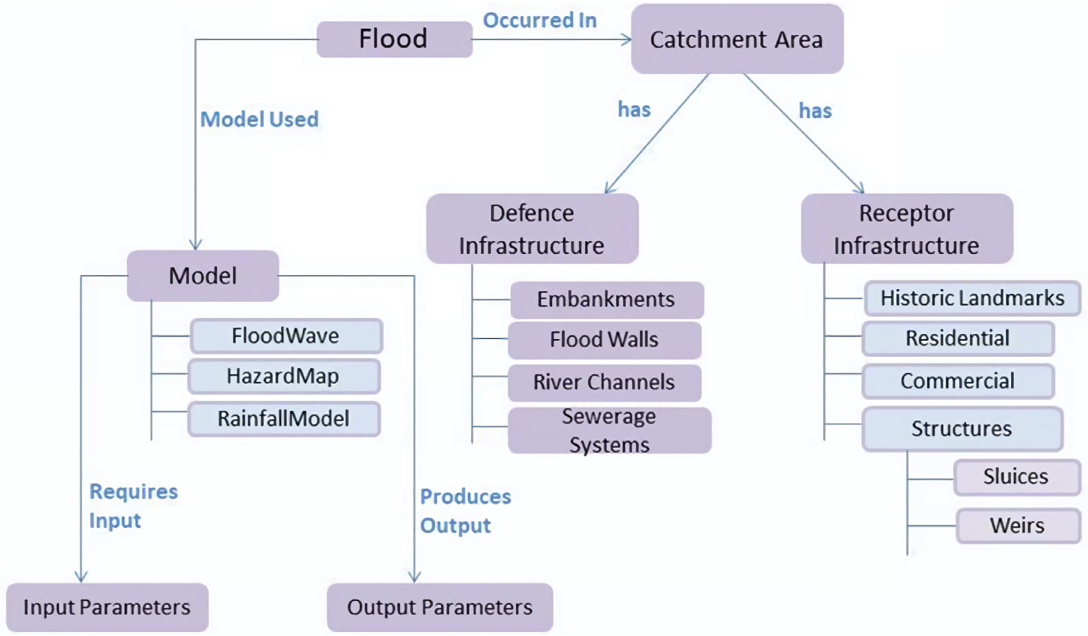


Fig. 3. Snapshot of EIA Ontology extended for the flood domain.

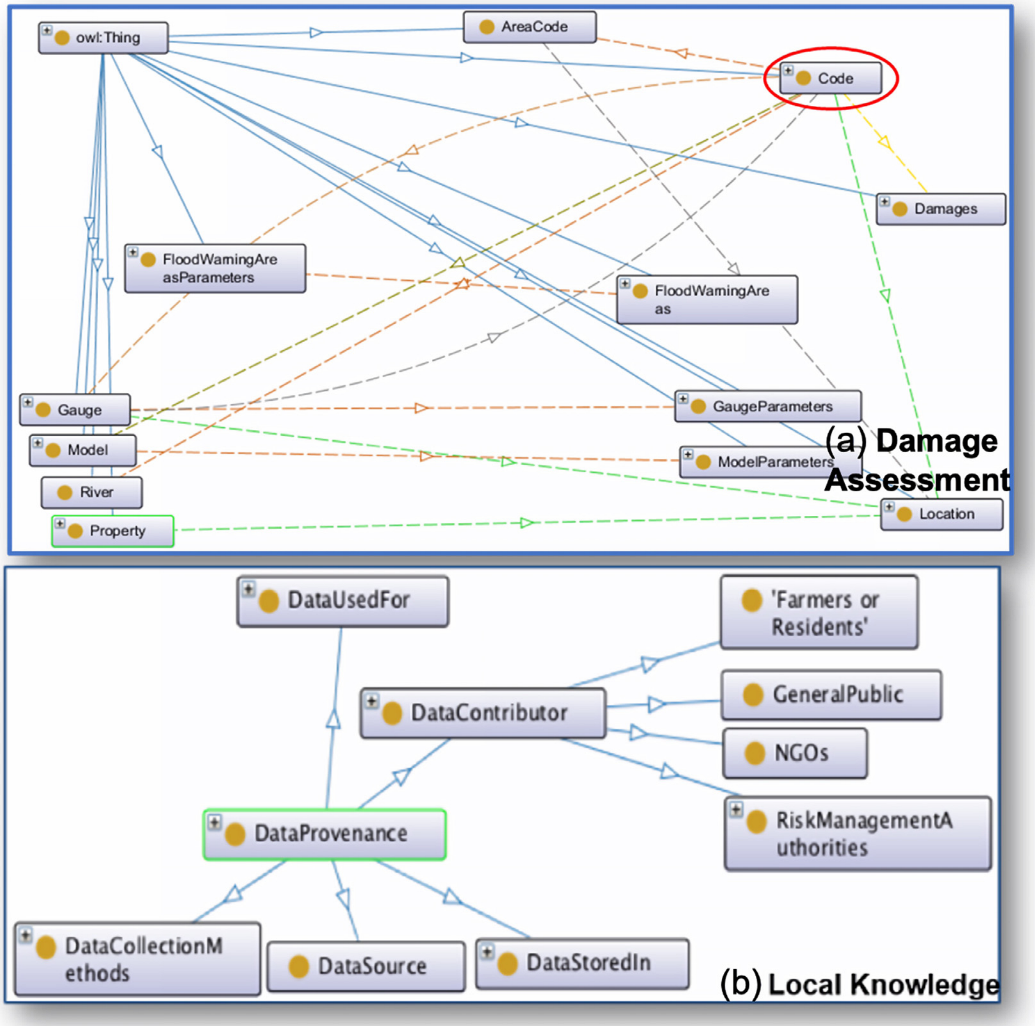


Fig. 4. Data Ontologies - Damage Assessment and Local Knowledge

a number of issues that need to be dealt with first before working with such data - such as some field values are left blank; some field values of type *string* may consist of multiple sub-strings, and may require aggregating or splitting as per the requirements of the application; some field labels contain special symbols, which are non-ASCII values (e.g. EstimatedDamage**£**); some field values contain both lowercase as well as uppercase characters and may require a standardised way to represent them. Whilst we have written a Python script to take care of all these irregularities, it is worth mentioning that there are software tools available for data cleaning/preparation, such as Talend Talend (2020), OpenRefine OpenRefine (2021) and OntoRefine GraphDB (2020). For instance, using Talend, we can even create data subsets if we have big data sets, and then semantically convert each subset separately.

5.3.2. Using the ontology to semantically enrich data

The domain ontology defines the domain concepts, and enables to stitch things like the provenance of information to the concepts involved, which is called the metadata. Fig. 5(a) shows the kind of metadata created around the *'River'* concept - such as cause of hazard for the river bursting its banks; the flood warning areas around the river etc. Such enrichment not only helps in enabling a deeper analysis of the data gathered about the concept but also helps to identify missing data gaps required for a deeper analysis. For example, from Fig. 5(a), one axiom says: *River hasCharacterizingIndicator some WasteWaterServices*. This points to the fact that if we have data about waste water services, affiliated with a particular river, it can help to give more insight on the actions to be taken around this river.

5.3.3. Using the ontology to integrate data

Fig. 2 illustrates how different data sets can be integrated through the data ontology and semantically enriched with metadata from the domain ontology. The example shown in Fig. 2 demonstrates a data ontology where '*Station*' is a term used to represent a gauge by a river, and hence can abstract heterogeneous representations of the 'Gauge' information found in the data sets. For example, '*GaugeWharfe*', *'GaugeRiverWharfe'*, etc. are all considered to be of type *'Station'* through such abstraction. Furthermore, the *'Station'* concept, being defined the same as the *'Gauge'* concept from the domain ontology (as shown in Fig. 2), inherits all the metadata properties of this domain ontology concept. This abstraction can also help resolve heterogeneities arising in data ontologies as well. For example, '*Station*' concept from one data ontology and '*Base*' concept from another data ontology, both representing gauge measurements, can be further abstracted to the *'Gauge'* concept in the domain ontology, thus ensuring integration across all the different data sets (as shown in Fig. 2). Moreover, the domain ontology represents the different concepts affiliated to the flood domain such as: *a River has a Gauge; a River has a Defence; a Defence is next to a Property; a River has a Location*. Given these relationships among the flood concepts ('River', 'Defence', 'Property', etc.), the semantically-enriched data also inherit these relationship properties. Hence, if there are data about other concepts such as '*Defence*' or '*Property*', they can be represented similarly through data ontologies and integrated with the existing data sets and richer query sets can be formulated.

5.4. Semantic enrichment of unstructured data

5.4.1. Data transformation for unstructured data

Unlike structured data, the unstructured data in the Section19 report consists of sentences, tables and figures. The approach elicited here shows how we can extract information from such a document through the use of NLP techniques. Fig. 6(a) gives an extract of a page from the Section19 report.

5.4.2. Approach adopted to extract data from Section19 report

Fig. 6 illustrates the steps followed for the data transformation of one page from the Section19 report.

1. Slicing the PDF The first step is to slice the PDF file into individual pages. This is done owing to the size of the report which is around 56 pages. The idea is to treat each page individually, and each page is named according to the page number (e.g. **page12.pdf**).

2. Converting one PDF page into text The next step is to convert a particular page into text form (e.g. **page12.txt**). This step enables the user to browse which page is required for data extraction; hence, converting only required pages into text form. Here, it is worth mentioning that the user can also remove any unwanted sentences or text from the file. Moreover, one detail that is added in this text file is a header at the start of each page. Related pages will have the same header, which denotes the topic of interest represented by this page, and helps to provide useful context during data querying. For example, all pages under a section like *'River Wharfe'* will bear the header *'River Wharfe'*.

3. Extracting information (Subject-Predicate-Object (SPO) from the text using Natural Language Processing (NLP) techniques) At this stage, NLP techniques have been used to identify nouns and verbs from the text page (e.g. **page12.txt**). A python parser has been written, making use of the nltk library NLTK (2020), to identify nouns and verbs and to split the sentences into three parts similar to a triple (Subject-Predicate-Object). The python parser identifies the verbs, and eventually identifies the phrase preceding the verb as the subject, and the phrase succeeding the verb as the object. In this way, a sentence is rendered as an **SPO**. The python parser extracts the SPOs from the sentences from each page and saves them in a different text file (**e.g. page12-SPO.txt**). But at this stage, the resulting page (**e.g. page12-SPO.txt**) is checked for redundancy. This action is required since the nltk library recognises 'every' verb as a verb, and if there is a verb within a subject phrase or object phrase, it is also recognised as a verb, and obviously, the prefix and suffix of this verb will be automatically classified as subject and object respectively. This step is semi-automated as it requires the user to browse through the list of SPOs per text file and remove those that do not make any sense. This is not such a tedious process as the text file represents a single page from the report. This step can be classified as a data cleaning process.

4. Converting the SPOs into RDF triples Once the SPOs have been formulated, they can then be semantically converted into an appropriate RDF triple form. In order to maintain the uniformity between the SPO and the RDF data, rdf constructs from the rdf vocabulary RDFSchema (2014) have been used to semantically convert the unstructured text. One such example is shown below:

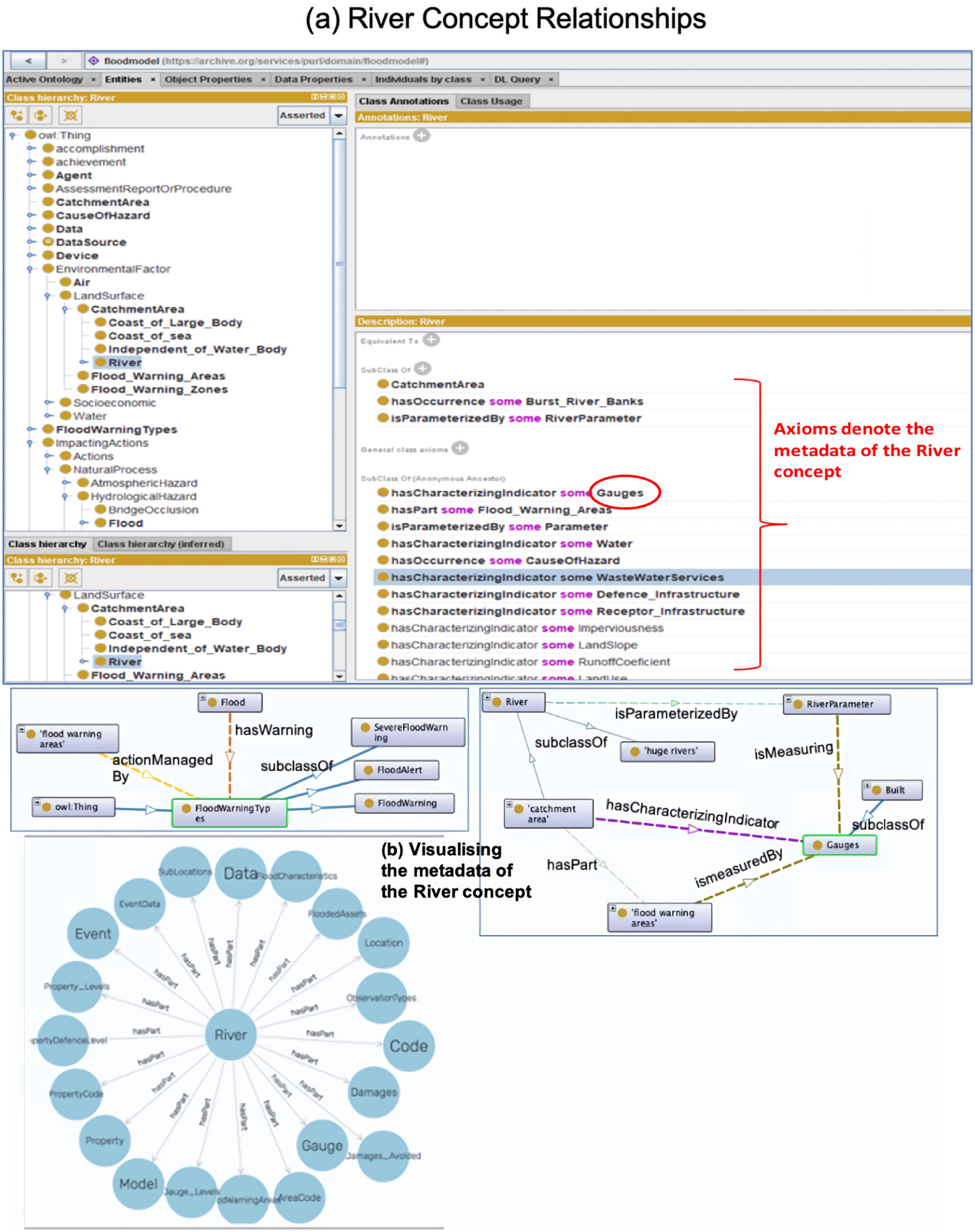
"**rdf:subject**" : "River Wharfe";
"**rdf:predicate**" : "has";
"**rdf:object**" : "GaugeName:Addingham Peak stage:2 Time:07:30 Date:26 Dec RankInRecord:5 RecordLength:43 CurrentorPreviousHighest:2.541 Jan 1982".

5.5. Loading and Querying the data from a semantic data store

Once the triples have been created, they are ready to be loaded onto a semantic data store for querying. The semantic augmentation has been carried out using a Python script which converts the data into RDF form and stores them in relevant JSON-LD files. These JSON-LD files can then be loaded onto the semantic store. The semantic data store used here is GraphDB (GraphDB, 2020) and the semantic query language used to query the semantic data is SPARQL (Sparql, 2008). The next section gives examples of queries executed from this unified model, and provides an evaluation of the GraphDB tool.

6. Evaluation

This section evaluates the linked data model qualitatively through some use cases, and also presents a quantitative analysis based on running the SPARQL queries on GraphDB. The benefit of using modular ontologies has also been highlighted in this section.



owl:Thing
- accomplishment
- achievement
- Agent
- AssessmentReportOrProcedure
- CatchmentArea
- CauseOfHazard
- Data
- DataSource
- Device
- EnvironmentalFactor
  - Air
  - LandSurface
    - CatchmentArea
      - Coast_of_Large_Body
      - Coast_of_sea
      - Independent_of_Water_Body
      - River
    - Flood_Warning_Areas
    - Flood_Warning_Zones
  - Socioeconomic
  - Water
- FloodWarningTypes
- ImpactingActions
  - Actions
  - NaturalProcess
    - AtmosphericHazard
    - HydrologicalHazard
      - BridgeOcclusion
      - Flood

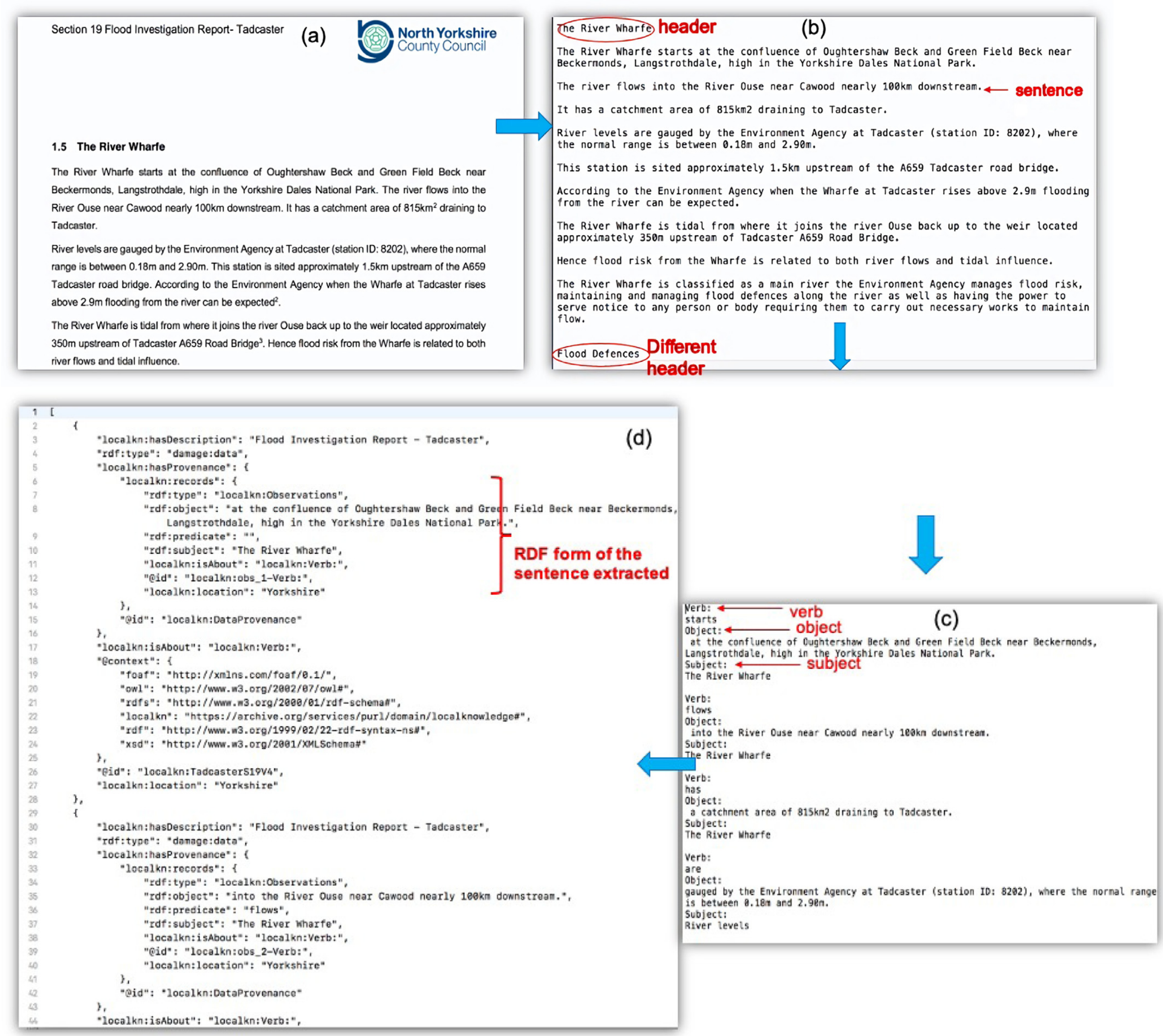
SubClass Of
- CatchmentArea
- hasOccurrence some Burst_River_Banks
- isParameterizedBy some RiverParameter

SubClass Of (Anonymous Ancestor)
- hasCharacterizingIndicator some Gauges
- hasPart some Flood_Warning_Areas
- isParameterizedBy some Parameter
- hasCharacterizingIndicator some Water
- hasOccurrence some CauseOfHazard
- hasCharacterizingIndicator some WasteWaterServices
- hasCharacterizingIndicator some Defence_Infrastructure
- hasCharacterizingIndicator some Receptor_Infrastructure
- hasCharacterizingIndicator some Imperviousness
- hasCharacterizingIndicator some LandSlope
- hasCharacterizingIndicator some RunoffCoeficient
- hasCharacterizingIndicator some LandUse

Axioms denote the metadata of the River concept

(b) Visualising the metadata of the River concept

Fig. 5. River Concept.


(a)
North Yorkshire County Council
(b)
header
sentence
Different header
(c)
verb
object
subject
(d)
RDF form of the sentence extracted

Section 19 Flood Investigation Report- Tadcaster

1.5 The River Wharfe

The River Wharfe starts at the confluence of Oughtershaw Beck and Green Field Beck near Beckermonds, Langstrothdale, high in the Yorkshire Dales National Park. The river flows into the River Ouse near Cawood nearly 100km downstream. It has a catchment area of 815km^2 draining to Tadcaster.

River levels are gauged by the Environment Agency at Tadcaster (station ID: 8202), where the normal range is between 0.18m and 2.90m. This station is sited approximately 1.5km upstream of the A659 Tadcaster road bridge. According to the Environment Agency when the Wharfe at Tadcaster rises above 2.9m flooding from the river can be expected[2].

The River Wharfe is tidal from where it joins the river Ouse back up to the weir located approximately 350m upstream of Tadcaster A659 Road Bridge[3]. Hence flood risk from the Wharfe is related to both river flows and tidal influence.

```
The River Wharfe

The River Wharfe starts at the confluence of Oughtershaw Beck and Green Field Beck near
Beckermonds, Langstrothdale, high in the Yorkshire Dales National Park.

The river flows into the River Ouse near Cawood nearly 100km downstream.

It has a catchment area of 815km2 draining to Tadcaster.

River levels are gauged by the Environment Agency at Tadcaster (station ID: 8202), where
the normal range is between 0.18m and 2.90m.

This station is sited approximately 1.5km upstream of the A659 Tadcaster road bridge.

According to the Environment Agency when the Wharfe at Tadcaster rises above 2.9m flooding
from the river can be expected.

The River Wharfe is tidal from where it joins the river Ouse back up to the weir located
approximately 350m upstream of Tadcaster A659 Road Bridge.

Hence flood risk from the Wharfe is related to both river flows and tidal influence.

The River Wharfe is classified as a main river the Environment Agency manages flood risk,
maintaining and managing flood defences along the river as well as having the power to
serve notice to any person or body requiring them to carry out necessary works to maintain
flow.

Flood Defences
```

```
Verb:
starts
Object:
 at the confluence of Oughtershaw Beck and Green Field Beck near Beckermonds,
Langstrothdale, high in the Yorkshire Dales National Park.
Subject:
The River Wharfe

Verb:
flows
Object:
 into the River Ouse near Cawood nearly 100km downstream.
Subject:
The River Wharfe

Verb:
has
Object:
 a catchment area of 815km2 draining to Tadcaster.
Subject:
The River Wharfe

Verb:
are
Object:
gauged by the Environment Agency at Tadcaster (station ID: 8202), where the normal range
is between 0.18m and 2.90m.
Subject:
River levels
```

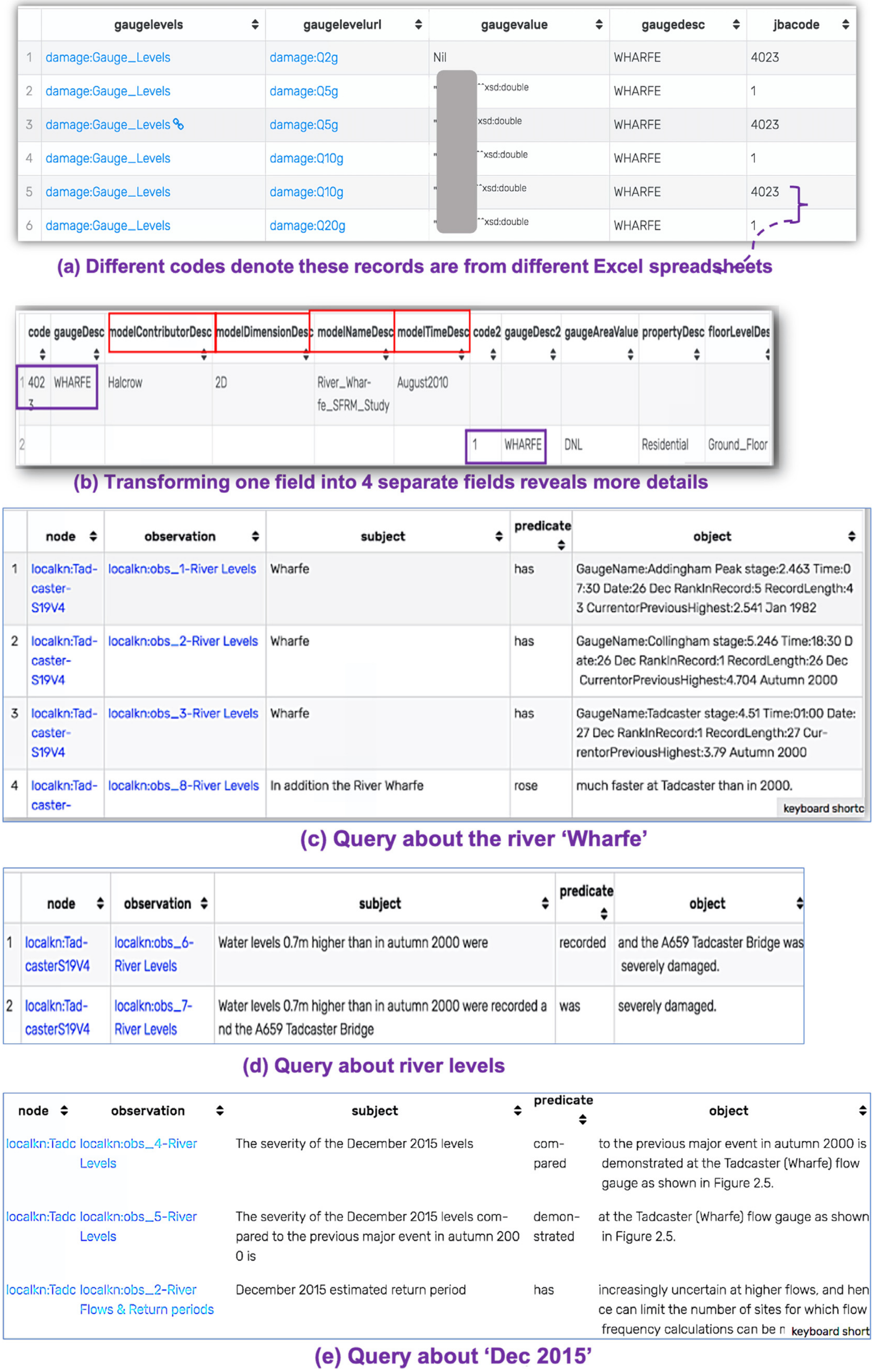
```
[
    {
        "localkn:hasDescription": "Flood Investigation Report - Tadcaster",
        "rdf:type": "damage:data",
        "localkn:hasProvenance": {
            "localkn:records": {
                "rdf:type": "localkn:Observations",
                "rdf:object": "at the confluence of Oughtershaw Beck and Green Field Beck near Beckermonds,
                    Langstrothdale, high in the Yorkshire Dales National Park.",
                "rdf:predicate": "",
                "rdf:subject": "The River Wharfe",
                "localkn:isAbout": "localkn:Verb:",
                "@id": "localkn:obs_1-Verb:",
                "localkn:location": "Yorkshire"
            },
            "@id": "localkn:DataProvenance"
        },
        "localkn:isAbout": "localkn:Verb:",
        "@context": {
            "foaf": "http://xmlns.com/foaf/0.1/",
            "owl": "http://www.w3.org/2002/07/owl#",
            "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
            "localkn": "https://archive.org/services/purl/domain/localknowledge#",
            "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
            "xsd": "http://www.w3.org/2001/XMLSchema#"
        },
        "@id": "localkn:TadcasterS19V4",
        "localkn:location": "Yorkshire"
    },
    {
        "localkn:hasDescription": "Flood Investigation Report - Tadcaster",
        "rdf:type": "damage:data",
        "localkn:hasProvenance": {
            "localkn:records": {
                "rdf:type": "localkn:Observations",
                "rdf:object": "into the River Ouse near Cawood nearly 100km downstream.",
                "rdf:predicate": "flows",
                "rdf:subject": "The River Wharfe",
                "localkn:isAbout": "localkn:Verb:",
                "@id": "localkn:obs_2-Verb:",
                "localkn:location": "Yorkshire"
            },
            "@id": "localkn:DataProvenance"
        },
        "localkn:isAbout": "localkn:Verb:",
```

Fig. 6. Semantic enrichment of unstructured data.

6.1. Qualitative analysis of the linked data model

The linked data model was set up by integrating different data sets together on GraphDB. These data sets were represented principally by the information extracted from the Section19 report, the Communities at Risk data set and the metadata about flood concepts provided through the flood ontologies. Not to mention that we can bring in other flood-related data sets in this model if available. The linked data model was evaluated to see if we can query for information from the individual datasets and also to see if we can query from multiple datasets behaving as one unified model. Therefore, the aim is to show how the information spread across the spreadsheets can be brought together, how the information extracted from the Section19 report can be queried, how the metadata of the 'raw' data can be queried from the flood ontology and how we can query for finer-grained details embedded within the Communities at Risk data set that have emerged as a result of applying data transformation to this set. Most essentially, we also show how the linked data model can be queried as an integrated model, pulling data sitting across structured and unstructured data sets - the Section19 report and the Communities at Risk data set - in order to reveal combined details as opposed to querying them individually. These queries were executed using a SPARQL query facility provided through the GraphDB tool and have been presented below:

- Integration of structured data provided through the Excel spreadsheets:
  Fig. 7 (a) shows the results of a query across different Excel spreadsheets for details about a gauge station near a river named “Wharfe”. The different codes highlighted in the figure denote different data records pulled out from the spreadsheets.
- Data transformation on the structured data:
  Fig. 7 (b) highlights 4 fields, which are a result of splitting one single field from one Excel spreadsheet. The aim behind this data transformation is to reveal some crucial details that are relevant to the flood scientists. For instance, the example shown in this figure shows details such as: who created a flood model, what is the dimension of the



	gaugelevels	gaugelevelurl	gaugevalue	gaugedesc	jbacode
1	damage:Gauge_Levels	damage:Q2g	Nil	WHARFE	4023
2	damage:Gauge_Levels	damage:Q5g	" [redacted] ^^xsd:double	WHARFE	1
3	damage:Gauge_Levels	damage:Q5g	" [redacted] xsd:double	WHARFE	4023
4	damage:Gauge_Levels	damage:Q10g	" [redacted] ^^xsd:double	WHARFE	1
5	damage:Gauge_Levels	damage:Q10g	" [redacted] ^^xsd:double	WHARFE	4023
6	damage:Gauge_Levels	damage:Q20g	" [redacted] ^^xsd:double	WHARFE	1

(a) Different codes denote these records are from different Excel spreadsheets

	code	gaugeDesc	modelContributorDesc	modelDimensionDesc	modelNameDesc	modelTimeDesc	code2	gaugeDesc2	gaugeAreaValue	propertyDesc	floorLevelDes
1	4023	WHARFE	Halcrow	2D	River_Wharfe_SFRM_Study	August2010					
2							1	WHARFE	DNL	Residential	Ground_Floor

(b) Transforming one field into 4 separate fields reveals more details

	node	observation	subject	predicate	object
1	localkn:TadcasterS19V4	localkn:obs_1-River Levels	Wharfe	has	GaugeName:Addingham Peak stage:2.463 Time:07:30 Date:26 Dec RankInRecord:5 RecordLength:43 CurrentorPreviousHighest:2.541 Jan 1982
2	localkn:TadcasterS19V4	localkn:obs_2-River Levels	Wharfe	has	GaugeName:Collingham stage:5.246 Time:18:30 Date:26 Dec RankInRecord:1 RecordLength:26 Dec CurrentorPreviousHighest:4.704 Autumn 2000
3	localkn:TadcasterS19V4	localkn:obs_3-River Levels	Wharfe	has	GaugeName:Tadcaster stage:4.51 Time:01:00 Date:27 Dec RankInRecord:1 RecordLength:27 CurrentorPreviousHighest:3.79 Autumn 2000
4	localkn:Tadcaster-	localkn:obs_8-River Levels	In addition the River Wharfe	rose	much faster at Tadcaster than in 2000.

keyboard shortc

(c) Query about the river 'Wharfe'

	node	observation	subject	predicate	object
1	localkn:TadcasterS19V4	localkn:obs_6-River Levels	Water levels 0.7m higher than in autumn 2000 were	recorded	and the A659 Tadcaster Bridge was severely damaged.
2	localkn:TadcasterS19V4	localkn:obs_7-River Levels	Water levels 0.7m higher than in autumn 2000 were recorded and the A659 Tadcaster Bridge	was	severely damaged.

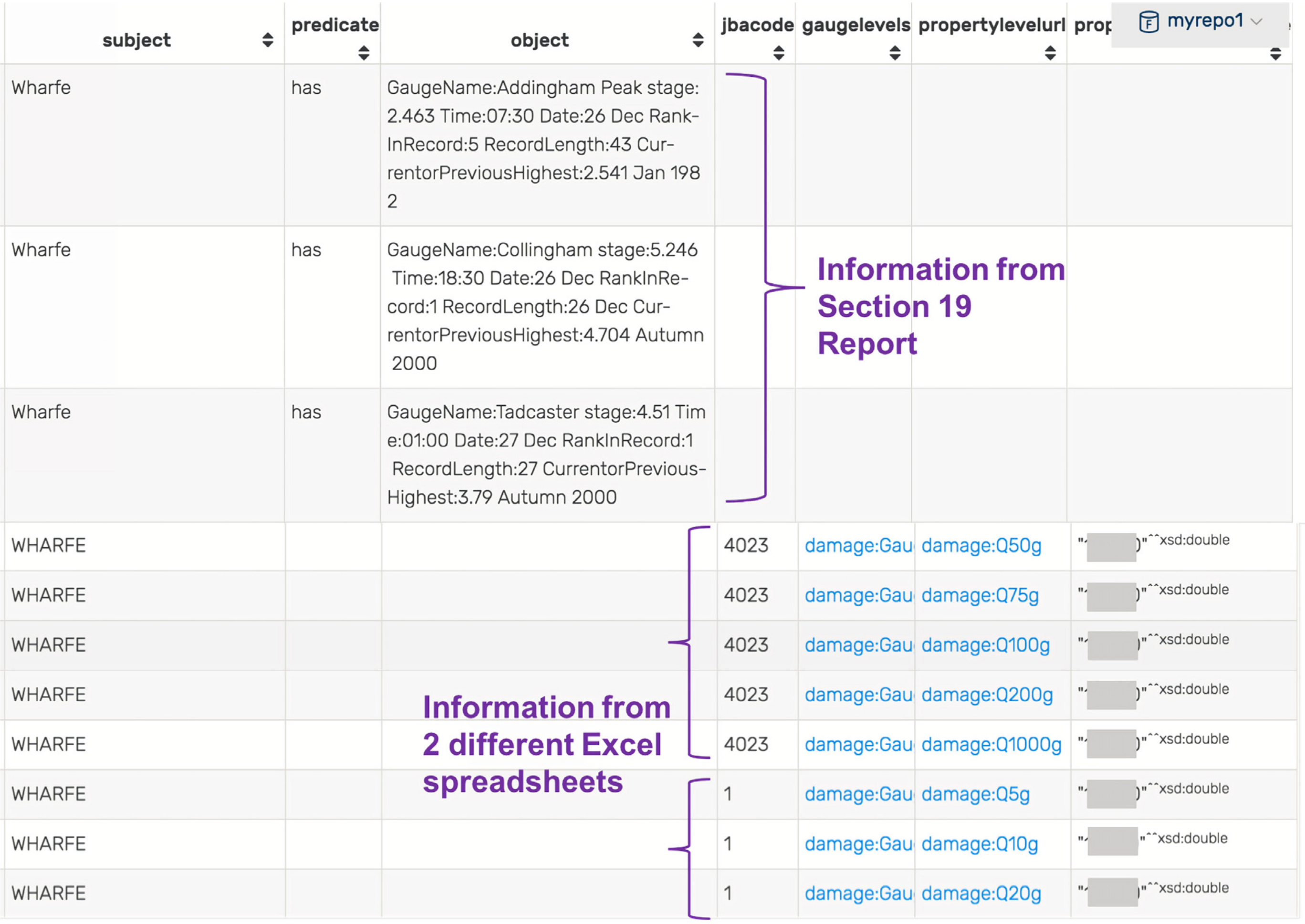
(d) Query about river levels

node	observation	subject	predicate	object
localkn:Tadc	localkn:obs_4-River Levels	The severity of the December 2015 levels	compared	to the previous major event in autumn 2000 is demonstrated at the Tadcaster (Wharfe) flow gauge as shown in Figure 2.5.
localkn:Tadc	localkn:obs_5-River Levels	The severity of the December 2015 levels compared to the previous major event in autumn 2000 is	demonstrated	at the Tadcaster (Wharfe) flow gauge as shown in Figure 2.5.
localkn:Tadc	localkn:obs_2-River Flows & Return periods	December 2015 estimated return period	has	increasingly uncertain at higher flows, and hence can limit the number of sites for which flow frequency calculations can be n

keyboard short

(e) Query about 'Dec 2015'

Fig. 7. Querying data.



subject	predicate	object	jbacode	gaugelevels	propertylevelurl	prop
Wharfe	has	GaugeName:Addingham Peak stage:2.463 Time:07:30 Date:26 Dec RankInRecord:5 RecordLength:43 CurrentorPreviousHighest:2.541 Jan 1982				
Wharfe	has	GaugeName:Collingham stage:5.246 Time:18:30 Date:26 Dec RankInRecord:1 RecordLength:26 Dec CurrentorPreviousHighest:4.704 Autumn 2000				
Wharfe	has	GaugeName:Tadcaster stage:4.51 Time:01:00 Date:27 Dec RankInRecord:1 RecordLength:27 CurrentorPreviousHighest:3.79 Autumn 2000				
WHARFE			4023	damage:Gau	damage:Q50g	"[redacted])"^^xsd:double
WHARFE			4023	damage:Gau	damage:Q75g	"[redacted])"^^xsd:double
WHARFE			4023	damage:Gau	damage:Q100g	"[redacted])"^^xsd:double
WHARFE			4023	damage:Gau	damage:Q200g	"[redacted])"^^xsd:double
WHARFE			4023	damage:Gau	damage:Q1000g	"[redacted])"^^xsd:double
WHARFE			1	damage:Gau	damage:Q5g	"[redacted])"^^xsd:double
WHARFE			1	damage:Gau	damage:Q10g	"[redacted]"^^xsd:double
WHARFE			1	damage:Gau	damage:Q20g	"[redacted])"^^xsd:double

Information from Section 19 Report

Information from 2 different Excel spreadsheets

Fig. 8. Integrating Excel data with Section19 report.

model, what is the name of the model, and when was the model created. The availability of more information through such data transformation not only reveals important information to the user, but also means that more data can be integrated given their availability.

- Querying an unstructured document:
  Fig. 7 (c-e) shows the results of querying the Section19 report. Information pertaining to the *river 'Wharfe'* (Fig. 7 (c)), *water levels* (Fig. 7 (d)) and the time period *Dec 2015* (Fig. 7 (e)) are shown as queries from the Section19 report. This flexibility of querying such a highly complex document emphasizes the benefits of using a semantic approach to make such information available.
- Semantic enrichment of data with metadata from the flood domain ontology:
  Fig. 5(b) shows the metadata of the river concept, and reflects how a particular concept is related to other concepts through the ontology. Querying such metadata not only helps the scientists to see how the data has been enriched but also enables them to identify data gaps in their analysis.
- Integration of structured data (Excel spreadsheets) together with unstructured data (Section19 report):
  Fig. 8 shows the results of integrating structured data and unstructured data. This query demonstrates the power of creating a unified data model over data which would have otherwise been silo-ed, and such integration can enable the flood scientists to gain a better insight about a given location/river. This particular use case shows observed measurements on the river "Wharfe" but also shows added information from the Section19 report.

Table 1
Triples queried.

Query Type	Triples Queried
Query river "Wharfe" from Excel sheets	174
Query all observations from Section19 document	2043
Query by Gauge	39
Query for a given gauge code	2
Query Property Return Periods for river "Wharfe"	320
Query river "Wharfe" across Excel sheets and Section19 document	213

6.2. *Quantitative analysis of the linked data model*

This section gives an overview of the number of triples formed during the integration process, and the time taken to run every query. Table 1 gives a glimpse of the number of triples produced from executing different queries against the Excel files and the Section19 document.

Although the number of triples output for each query varies in a range of 2 to 2043, we note that the time taken for running the queries is trivial (0.1-0.2 sec). The GrapdhDB query facility (SPARQL) may be efficient, but also the GraphDB version used is one of a standalone server, implying that the triple store is running on the same machine where we query the triples. Another observation is that although we converted only 11 pages (out of 56) of data into semantic format, there were a lot of triples generated (2043). This is due to the fact that textual data is more compact, and although one page of data seems trivial compared to a long list of records from an Excel sheet, one page can generate many triples owing to the high number of sentences it may contain. On the other hand, there was a lot of data regarding properties that was also converted (4000 records), but the relatively small number of triples resulting from querying any property data was due to the fact that the properties may not be affiliated to the gauge or river data that was loaded. To load the whole property data (around 19800 records), we would require more computing resources in terms of loading the GraphDB on a server machine or on a cloud-based version to accommodate the high quantity of triples generated.

6.3. Benefit of modular ontologies

The data ontology can be imported into the domain ontology in order to reflect the existence of the schematic data in the domain context. For example, the *'Input Parameters'* (see Fig. 3) is a general concept in the flood domain ontology, but it can get contextualised through the data ontology. Hence, the idea behind keeping the ontologies modular is to enable the reuse of the domain ontology with some other data ontology; and likewise reuse the data ontology, if required, with some other domain ontology. This way, we can exert a better control over the ontologies regarding their maintenance and update. We do not need to disrupt the entire flood ontological model if we need to add a new domain concept or change the data schema. One reason behind keeping ontologies modular is due to the high heterogeneity arising in the data. Using only a global ontology (see Section 3) would imply constantly updating the ontology with new concepts to accommodate new data, thus making the ontology more bulky and unmanageable.

7. Discussion

This paper shows an approach to extract numerical as well as textual information from an unstructured source and how to bring this together with information from a structured source. Regarding the unstructured data, we have extracted information from sentences and tables from a PDF file but the approach highlighted can be applied to other types of textual sources. JBA Trust/JBA Consulting are enhancing this prototype model by looking at ways to extract further information from the flood reports, either from text or images, and integrate them with other datasets for further analysis. We also envisage to use Natural Language Processing techniques to extract relevant pieces of information from textual data that may be found within the subject or object entities.

The modular ontological approach adopted in this paper enables a better maintenance of the domain and data ontologies. It allows for greater flexibility in terms of updating only the data ontologies in order to accommodate new data sources without affecting the domain ontology. Since the domain ontology models higher level concepts, it remains unchanged even if the data parameters are changed. Depending on the nature of the application, it is sometimes better to create a new ontology rather than reuse an existing one which can be bulky. The ontology design however requires the input of a domain expert and also of a developer with some experience in ontologies. Ontologies can also be used to make inference of new knowledge based on their reasoning capabilities. For example, a reasoning can indicate whether there is a danger of water pollution in a flooded area if the water has been found to contain excess of nitrogen and phosphorus. We intend to infer information through ontological reasoning in the future.

GraphDB has been used as the triple store as it can support millions of triples, and is also now available as a cloud version to host data. The data we used was for a prototype implementation; however, on a larger scale, triple stores such as GraphDB can deliver highly scalable solutions. Regarding the SPARQL queries, environmental scientists will need the assistance of a developer who has the knowledge of using the SPARQL language. However, further work can be done around writing queries in simple English or developing a visual query interface, and mapping them automatically onto their corresponding SPARQL query. There has been some work done around this such as Ferré (2013). Such an approach can enable scientists to focus more on running their queries rather than figuring out how to formulate them in SPARQL. Moreover, the use of semantic techniques looks like a promising step towards preventing data from being silo-ed, and also helping scientists towards better decision-making process.

8. Conclusion

The need to reduce uncertainty in flood modelling and flood risk management more generally is imperative for better decision making and policy making. Flood scientists stipulate that bringing in a broader spectrum of parameters into the modelling process can help alleviate the uncertainty dilemma. Although flood risk management largely depends on measurement data, there is a need to bring in data from different dimensions in order to understand the unknown parameters in modelling. A huge plethora of information in flood reports can help bring better insight on flood risk analysis. However, their unstructured nature impose a barrier on the usage of the contained information for analytical purposes. Previously, the emphasis of data integration approaches was more on structured datasets, but there is now a gradual shift towards handling unstructured data sources as well especially with technological advances in Machine Learning and Natural Language Processing. Existing approaches allow the extraction of textual entities from text, but since we are dealing with both numerical and textual data in flood reports, extracting only textual entities is not sufficient. We need to extract numerical values from text without losing their contextual information.

Hence, this paper has demonstrated how we can extract numerical as well as textual data from a document whilst preserving their integrity and how we can bring them together with data from structured sources to form a unified model for richer querying. The linked data model shows that it is possible to bridge the gap between structured and unstructured sources, and hence support our vision for an enriched support for collaboration and decision-making in flood risk management. The roles of Semantic Web and Natural Language Processing techniques have been highlighted to enable such integration. This combination of technologies has successfully addressed our three research questions identified in the introduction. In particular, we have demonstrated how we can successfully achieve data integration including the incorporation of unstructured and structured sources and subsequently applied queries that draw on both sources of information. Our use of a hybrid ontological model of domain and data ontologies to integrate the heterogeneous data sources has also been particularly successful, most notably in facilitating the update and maintenance of the ontologies.

Our approach has been demonstrated and evaluated through a real world case study and one important area of future work would be to carry out further case studies drawing on different aspects of flood risk management. We also believe that the approaches advocated in this paper have broader applicability across other areas of environmental science and we are currently considering the use of Natural Language Processing and Machine Learning techniques to extract information on plants, observers and locations from historical archives relating to biodiversity.

Funding

This research was funded by Engineering and Physical Sciences Research Council Grant EP/P002285/1 (SeniorFellowship on the Role of Digital Technology in Understanding, Mitigating and Adapting to EnvironmentalChange).

Declaration of Competing Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript; or in the decision to publish the results.

Acknowledgment

We would also like to acknowledge Ross Towe for his insight on the Communities At Risk data set, and Graham Dean and Elizabeth Edwards for very helpful discussions towards the design of the flood domain ontology.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.envc.2021.100064

References

Araque, F., Salguero, A., Delgado, C., 2008. Ontology based framework for data integration. WSEAS Trans. Inf. Sci. Appl. 5, 953–962.

Arens, Y., Knoblock, C., Shen, W.m., 1996. Query reformulation for dynamic information integration. J. Intell. Inf. Syst. 6, 99–130.

Astera, 2020. Astera. Retrieved from https://www.astera.com/

Atemezing, G., 2015. Semantic web methodologies, best practices and ontology engineering applied to internet of things. In: Proceedings of the Conference IEEE World Forum - Internet Things, pp. 412–417.

Baldassarre, G.D., Brandimarte, L., Beven, K., 2016. The seventh facet of uncertainty: wrong assumptions, unknowns and surprises in the dynamics of human–water systems. Hydrol. Sci. J. 61, 1748–1758.

Bernadette, F.L., Zoubida, K., Bouzeghoub, M., Assia, S., 2002. Heterogeneous data source integration and evolution. In: Proceedings of the 13th International Conference Database Expert Systems Applications, pp. 751–757.

Beven, K., Smith, P., 2015. Concepts of information content and likelihood in parameter calibration for hydrological simulation models. J. Hydrol. Eng. 20 (1), A4014010.

Blair, G., Beven, K., Lamb, R., Bassett, R., Cauwenberghs, K., Hankin, B., Dean, G., Hunter, N., Edwards, L., Nundloll, V., Samreen, F., Simm, W., Towe, R., 2019. Models of everywhere revisited: a technological perspective. Environ. Modell. Softw. 122, 104521.

Collet, C., Huhns, M.N., Shen, W.M., 1991. Resource integration using a large knowledge base in Carnot. Computer. 24 (12), 55–62.

Cruz, I.F., Xiao, H., 2003. Using a layered approach for interoperability on the semantic web. In: Proceedings of the 4th International Conference Web Information Systems Engineering WISE 2003, pp. 221–231.

d'Aquin, M., Noy, N.F., 2012. Review: where to publish and find ontologies? A survey of ontology libraries. Web Semant. 11, 96–111.

Demir, I., Krajewski, W.F., 2013. Towards an integrated flood information system: centralized data access, analysis, and visualization. Environ. Model. Softw. 50, 77–84.

Dolbear, C., Goodwin, J., Mizen, H., Ritchie, J., 2005. Semantic interoperability between topographic data and a flood defence ontology.

Environment-Agency, 2020. National flood risk assessment 2 (NAFRA2).Retrieved from https://environment-analyst.com/uk/105640/environment-agency-awards-jacobs-8m-nafra2-contract.

Environment-Agency, 2021. Environment agency (UK). Retrieved from https://www.gov.uk/government/organisations/environment-agency.

Environment-Agency-Cost, 2018. Estimating the economic costs of the 2015 to 2016 winter floods.Retrieved from https://www.gov.uk/government/publications/floods-of-winter-2015-to-2016-estimating-the-costs.

Fernández-López, M., Gómez-Pérez, A., Suárez-Figueroa, M.C., 2013. Methodological guidelines for reusing general ontologies. Data Knowl. Eng. 86, 242–275.

Ferrario, M.A., Simm, W., Newman, P., Forshaw, S., Whittle, J., 2014. Software engineering for 'social good': integrating action research, participatory design, and agile development. In: Proceedings of the 36th International Conference Softw. Eng., pp. 520–523.

Ferré, S., 2013. squall2sparql: a translator from controlled english to full sparql 1.1. In: Proceedings of the CEUR Workshop, 1179.

FloodReport, 2021. Lancashire flood report. Retrieved from http://www.lancashire.gov.uk/council/performance-inspections-reviews/environmental/flood-investigation-report.

Garrido, J., Requena, I., 2011. Proposal of ontology for environmental impact assessment: an application with knowledge mobilization. Expert Syst. Appl. 38, 2462–2472.

Goasdoué, F., Reynaud, C., 1999. Modeling information sources for information integration. In: Proceedings of the International Conference Knowl. Acquis., pp. 121–138.

GraphDB, 2020. Graphdb semantic graph database. Retrieved from http://graphdb.ontotext.com/free/loading-data-using-ontorefine.html.

Gruber, T.R., 1993. A translation approach to portable ontology specifications. Knowl. Acquis. 5, 199–220.

He, Z., Tao, C., Bian, J., Dumontier, M., Hogan, W., 2017. Semantics-powered healthcare engineering and data analytics. J. Healthc. Eng. 2017, 1–3.

Horecki, K., Mazurkiewicz, J., 2015. Natural language processing methods used for automatic prediction mechanism of related phenomenon. In: Proceedings of the 20th International Conference Artif. Intell. Soft Comput., 9120, pp. 13–24.

IBM, 2020. IBM technology products. Retrieved from https://www.ibm.com.

Ison, J., Kalaš, M., Jonassen, I., Bolser, D., Uludag, M., McWilliam, H., Malone, J., Lopez, R., Pettifer, S., Rice, P., 2013. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics 29, 1325–1332.

JBAConsulting, 2021. JBA consulting. Retrieved from https://www.jbaconsulting.com/

JBATrust, 2021. JBA trust. Retrieved from https://www.jbatrust.org/.

McCallum, I., Liu, W., See, L., Mechler, R., Keating, A., Hochrainer-Stigler, S., Mochizuki, J., Fritz, S., Dugar, S., Arestegui, M., Szoenyi, M., Bayas, J.-C.L., Burek, P., French, A., Moorthy, I., 2016. Technologies to support community flood disaster risk reduction. Int. J. Disaster Risk Sci. 7, 198–204.

Mena, E., Illarramendi, A., Kashyap, V., Sheth, A., 2000. Observer: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. Distrib. Parallel Dat. 8, 223–271.

MontoriF., Bedogni, L., Bononi, L., 2016. On the integration of heterogeneous data sources for the collaborative internet of things.. In: Proceedings of the IEEE 2nd International Forum Res. Technol. Soc. Ind. Leveraging a Better Tomorrow, RTSI, pp. 1–6.

NLTK, 2020. Natural language toolkit. Retrieved from https://www.nltk.org.

Ontotext - knowledge graph technology, 2021a. Ontotext - semantic data integration. Retrieved from https://www.ontotext.com/.

Ontotext, 2021b. Ontotext-knowledge graph technology. Retrieved from https://www.ontotext.com/knowledgehub/fundamentals/semantic-data-integration.

OpenRefine, 2021. Openrefine data cleaning tool. Retrieved from http://openrefine.org/.

Protege, 2020. Protege-ontology editor and framework. Retrieved from http://protege.stanford.edu/.

RDF, 2014. Resource description framework. Retrieved from https://www.w3.org/RDF/.

RDFSchema, 2014. Resource description framework schema. Retrieved from https://www.w3.org/TR/rdf-schema/.

Romanov, A., Konstantin, L., Kozlova, E., 2019. Application of natural language processing algorithms to the task of automatic classification of russian scientific texts. Data Sci. J. 18, 37.

SemanticWeb, 2021. Semantic web. Retrieved from https://www.w3.org/standards/semanticweb/.

Smith, L., Liang, Q., James, P., Lin, W., 2017. Assessing the utility of social media as a data source for flood risk management using a real-time modelling framework. J. Flood Risk Manag. 10, 370–380.

Sparql, 2008. Sparql query facility. Retrieved from https://www.w3.org/TR/rdf-sparql-query/.

Sun, K., Zhu, Y., Pan, P., Hou, Z., Wang, D., Li, W., Song, J., 2019. Geospatial data ontology: the semantic foundation of geospatial data integration and sharing. Big Earth Data 3, 269–296.

Talend, 2020. Talend data preparation tool. Retrieved from https://www.talend.com/products/data-preparation.

Towe, R., Dean, G., Edwards, L., Nundloll, V., Blair, G., Lamb, R., Hankin, B., Manson, S., 2020. Rethinking data-driven decision support in flood risk management for a big data age. J. Flood Risk Manag. 13, e12652.

Turki, H., Shafee, T., Hadj Taieb, M.A., Ben Aouicha, M., Vrandečić, D., Das, D., Hamdi, H., 2019. Wikidata: a large-scale collaborative ontological medical database. J. Biomed. Inform. 99, 103292.

Zhang, H., Guo, Y., Li, Q., George, T., Shenkman, E., Modave, F., Bian, J., 2018. An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. BMC Med. Inform. Decis. Mak. 18.

Ziegler, P., Dittrich, K.R., 2004. User-specific semantic integration of heterogeneous data: the sirup approach. In: Semantics of a Networked World. Semantics for Grid Databases, 3226, pp. 44–64.