Article:

Rosser, J., Jackson, M. and Leibovici, D. orcid.org/0000-0002-3903-2443 (2018) Full Meta Object profiling for flexible geoprocessing workflows. Transactions in GIS, 22 (5). pp. 1221-1237. ISSN 1361-1682

https://doi.org/10.1111/tgis.12460

Full Metadata Object profiling for flexible geoprocessing workflows

Julian F. Rosser, Mike Jackson and Didier G. Leibovici

University of Nottingham, Nottingham, UK

julian.rosser@nottingham.ac.uk

Full Metadata Object Profiling for flexible geoprocessing workflows

The design and running of complex geoprocessing workflows is an increasingly common geospatial modelling and analysis task. The Business Process Model and Notation (BPMN) standard, which provides a graphical representation of a workflow, allows stakeholders to discuss the scientific conceptual approach behind this modelling while also defining a machine-readable encoding in XML. Previous research has enabled the orchestration of Open Geospatial Consortium (OGC) Web Processing Services (WPS) with a BPMN workflow engine. However, the need for direct access to pre-defined data inputs and outputs results in a lack of flexibility during composition of the workflow and of efficiency during execution. This article develops metadata profiling approaches, described as two possible configurations, which enable workflow management at the meta-level through a coupling with a metadata catalogue. Specifically, a WPS profile and a BPMN profile are developed and tested using open-source components to achieve this coupling. A case study in the context of an event mapping task applied within a big data framework and based on analysis of the Global Database of Event Language and Tone (GDELT) database illustrates the two different architectures.

Keywords: workflow, metadata, catalogue services, web processing service, big data

1 Introduction

Geoprocessing workflows are a fundamental concept to the development of geospatial applications and products (Alonso & Hagen, 1997; De Giovanni et al., 2016; Hu, Wu, Zhong, Lv, & Yu, 2010; Nativi, Mazzetti, & Geller, 2013; Sun & Yue, 2010). Applications such as data conflation, quality assurance procedures and cartographic production are typical geospatial analysis tasks where a repository of geoprocessing transformations is used as a toolbox to compose a 'chain' of operations, and where reusing these chains as-is or with few modifications can often be needed. For example, standardised interfaces of legacy GIS analysis components may be exposed for creating these workflows e.g. Yue et al (2010) or new types of tests may be defined together to fulfil a particular experimental design e.g. Meek et al (2014). Both the toolboxes and available datasets can be shared within communities using interoperable e-infrastructures comprised of web services and may be described with metadata of the geoprocessing components available. Easy access to this metadata would be helpful as a support when composing and executing the workflows. Increasingly these workflows involve large data sets, distributed across different locations and computer systems, putting extra load on any geocomputational applications.

Service oriented approaches to computing offer a promising way to integrate different computer architectures, programming languages and processing needs required by geoprocessing workflows (Castronova, Goodall, & Elag, 2013; De Giovanni et al., 2016; Di, Shao, & Kang, 2013; Sheng et al., 2014; Sun & Yue, 2010). A Service Oriented Architecture (SOA) defines individual software components that provide data and functionality as Web services (Yang, Raskin, Goodchild, & Gahegan, 2010).Within the geospatial context, Open Geospatial Consortium web services (OWS) refers to services that are defined according OGC standards. OWS have some minimum required functionality that must be implemented (e.g. GetCapabilities) and respond to HTTP requests made from the clients. Although OWS are self-describing, and do include support for including metadata as part of their capabilities, the Catalogue Service for the Web (CSW) specification defines a standard for registering and locating metadata associated across multiple data or geoprocessing service instances (OGC, 2007b). This standardised metadata cataloguing enables client applications to efficiently identify and make use of the resources. Three ISO standards are relevant here for helping achieve this including ISO19115, ISO19119 and ISO19139. ISO19115 defines metadata that should be associated with a geographic resource such as its history, quality and intended use (ISO, 2003). ISO19119 is high-level standard that defines a hierarchical categorisation of six geospatial services including Workflow/Task services and Processing services, with the latter being further sub-divided into four further categories (ISO, 2005). ISO19139 is an XML schema that implements the ISO19115 standard and can be used to define metadata records (ISO, 2007). These records are inserted in, and retrieved from a CSW system.

When defined as services, chaining and orchestration of processes into workflows can be achieved. Two notable distinctions exist between service chaining and service orchestration. Service chaining is undertaken when multiple processes are combined to form a sequence or pipeline which creates a new service (Alameh, 2003). Web service orchestration can be defined as integrating the invocation of two or more services into a more complex workflow (Peltz, 2003). The orchestration can be a manual specification of outputs to other services, semi-automatic (through use of a configuration file), or automatic through the publication of capabilities between services (Kiehle, Greve, & Heier, 2007). Graphical environments for modelling are commonly used for scientific workflows and are attractive in a range of disciplines (De Giovanni et al., 2016; de Jesus, Walker, Grant, & Groom, 2012; Deelman, Gannon, Shields, & Taylor, 2009; Oinn et al., 2006).

In recent previous work, the use of BPMN for geoprocessing workflows has been adopted (Bigagli, Santoro, Mazzetti, & Nativi, 2015; Meek, Jackson, & Leibovici, 2016; Wiemann, 2016). BPMN is an Object Management Group (OMG) and ISO standard aimed at replacing flowchart diagrams and offers a graphical notation in association with the XML executable by the workflow engine. Bigagli et al. (2015) also made use of BPMN due its readily understandable graphical representation over Business Process Execution Language (BPEL). Meanwhile, Meek, Jackson, and Leibovici (2016) proposed orchestrating WPS through extending a Business Process

Modelling (BPM) platform which utilises BPMN workflow standard. The approach that was developed relied on direct management of the objects, data inputs and outputs, as well as the geoprocessing service defined as a customised workflow engine task.

In this paper, we describe the development and application of a profile-based architecture that couples a metadata catalogue with a workflow process modelling platform. The use of a self-contained BPMN file, which encodes the workflow, enables easy access to all the metadata associated with the geoprocessing, abstracts away the data and process objects from the workflow engine and delays use of the data objects during the workflow execution. This is achieved by designing and developing profiles for geoprocessing web services and BPMN based upon a metadata coupling, which we preliminarily described in short form in Rosser et al (2016) and extend here. Our contributions include:

- Design and development of two approaches for integrating metadata within a geoprocessing workflow,
- Illustration of the potential and usability of each solution with a comparison of the both approaches,
- Experimental demonstration of the proposed approaches within a big data geospatial workflow, thus enabling integration of different technology stacks,

The remainder of this paper is structured as follows: section 2 highlights related work; section 3 details the two profiling designs and architectural configurations as well discussing usability of the solutions provided; section 4 describes the implementation of the components; section 5 describes a case study for experimental deployment; finally, concluding remarks are made in section 6.

2 Related work on service orchestration

The focus of this work is on providing an interoperable environment that facilitates seamless re-use and sharing of scientific models composed as BPMN workflows, and relates to both web service orchestration and metadata management. In particular, the goal is also to be able to rapidly adapt existing open source tools to provide an environment capable of enabling further support of scientific modelling workflow management and execution, e.g. scenario testing and simulation, error propagation, and parallelisation. This section describes related literature in this field and highlights shortcomings with existing approaches regarding the ambitions stated above.

Integrating approaches for web service orchestration with OWS can present many difficulties for geospatial workflows. For example, the two systems typically do not use a common protocol. In particular, Web Service Description Language (WSDL) and Simple Object Access Protocol (SOAP) are used by web service orchestration engines but neither of these is well-adopted in OWS implementations, which instead tend to use Key-Value Pair (KVP) encoded in HTTP calls. Furthermore, Alameh (2003) identified that SOAP and WSDL are insufficient to describe geographic services which need to provide extra details about the spatial data such as capabilities and coverage. The difficulty in integrating OWS within the wider setting of web service orchestration has prompted various integration approaches. Version 1 of the WPS standard, the version adopted for this work, identifies three ways to chain services (OGC, 2007c). One approach is to use a BPEL engine to define and execute the workflow. This has previously been demonstrated as a mechanism for chaining WPS calls in relation to various applications (Brauner, Foerster, Schaeffer, & Baranski, 2009; Hobona, Fairbairn, Hiden, & James, 2010; Yu et al., 2012). However, this approach has been criticized by its technical complexity which may lead to non-domain users defining workflows (Bensmann, Alcacer-Labrador, Ziegenhagen, & Roosmann, 2014). Another option for orchestration is to wrap a sequence of WPS calls within another WPS

(Bielski, Gentilini, & Pappalardo, 2011; Eberle & Strobl, 2012). The third option mentioned in the WPS standard is to encode a chain of services within the execute query to form a cascading request. Version 2 of the WPS standard does not recommend any particular approaches for service chaining (Mueller & Pross, 2015).

With respect to providing intensive computing operations via WPS, Castronova (2013) developed a wrapper interface between WPS and an Open Modelling Interface (OpenMI) simulation framework and applied it to a hydrologic model case study. The WPS wrapper sits on the client machine and converts data into the OpenMI standard, and this enables a loose coupling of the scientific simulation model with OGC services. Use of self-describing packages of geoprocessing components have also been proposed as a mechanism to ease sharing of algorithms and scientific models that involve intensive analysis and large data sets (Müller, Bernard, & Kadner, 2013). Furthermore, the concept of a Geoprocessing Appstore presents one solution to help improve cataloguing and discovery of processes by acting as a central repository for the algorithm code together with machine-readable associated descriptions (Henzen, Brauner, Müller, Henzen, & Bernard, 2015). However, although the work adopts "Moving Code Packages" as a format for describing the algorithms, the catalogue does not provide metadata according to a standardised catalogue installation.

The inclusion of semantics has been shown to help with the design of workflows and with the documentation of the results of an analysis. Hobona et al (2007) propose a semantically-assisted system that enables a user to compose a workflow at an abstract level and then have a concrete implementation of it suggested based on a similarity score calculated using an ontology of the workflow components. Their system requires that the metadata for the workflow resource be tagged using Web Ontology Language (OWL) concept. Furthermore, utilising artificial intelligence path planning strategies alongside ontology descriptions of workflow components can also help with semi-automatic creation of geospatial workflows, and these descriptions may be encoded in ISO19115 documents in a CSW (Yue et al., 2009). Al-Areqi et al (2016) also describe annotating web-services with ontologies to enable automatic composition of workflows based on an initial sketch provided by the user.

With respect to aiding documentation of workflows, Yue et al (2010) developed a system for helping the tracking of metadata using semantic web technologies to capture provenance details within a service-oriented environment. Müller (2015) suggests that processes can be defined using a hierarchical profile of geoprocessing operations. For example, processes can have profiles at a conceptual level (i.e. what it does), which can be extended to a generic profile (i.e. how the operation is computed) which in turn can be extended to an implementation level (i.e. what data encoding is required and produced).

While progress has been made in using OGC services within a workflow environment, integrating metadata relating to the input data and processing implementations into the composition process has not been undertaken (Bigagli et al., 2015; Sheng et al., 2014), as proposed here.

3 Full Meta Objects Profiles & architecture

The aim of the work presented in this paper is to describe and test the required architecture to facilitate workflow composition using metadata records for the datasets and geoprocessing tasks used in the workflow. The principle followed for this architecture is to define the artefacts of the workflow, i.e. data and processes, from their metadata. The metadata links (exposed as URI strings) are managed either entirely by the workflow engine (BPMN implementation) or by the WPS behind each geoprocessing task. In practical deployment, this means that both data and processes are defined as a web-accessible metadata records (ISO 19139) and references to these records are embedded in the BPMN XML workflow definition.

Two approaches, named Full Meta Objects (FMO) profiles, for constructing and executing workflows comprised of metadata objects have previously been described (Rosser et al., 2016). Here we develop and test the architecture for an experimental processing scenario. Our proposal is the coupling of the workflow system with a metadata catalogue (see Figure 1). The configuration of system components is such that the workflow editor deals only with their metadata and does not need to understand the technicalities of the process inputs and outputs (such as the geospatial data formats). Similarly, depending on the type of profile architecture used, the workflow engine can also avoid needing to handle geospatial data and process entities (see section 4).

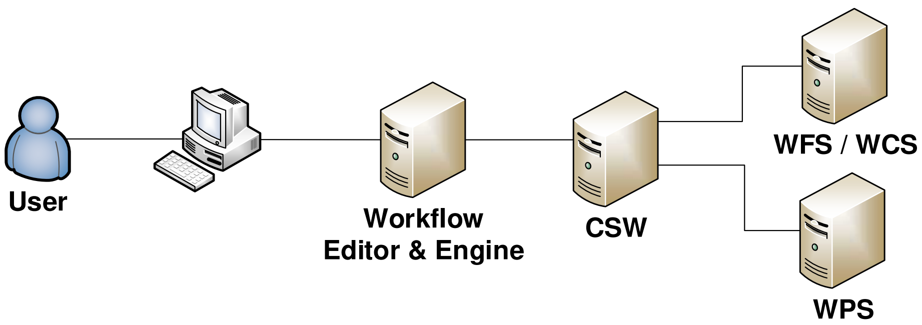

Figure 1. Overview of workflow composition using metadata objects. See Figure 2 and Figure 4 for detailed component diagrams illustrating options of system architecture and communication between components. The architecture uses open source software (described in further detail below) that implement the OGC and the BPMN standards. In particular, CSW: OGC Catalogue Services for the Web, WFS: OGC Web Feature Service and WPS: OGC Web Processing Service.

3.1 Web Processing Service profiling

The FMO WPS profile is a WPS that for each input accepts a metadata link to the metadata record of the dataset. Figure 2 illustrates the architecture between the components and Figure 3 shows the execution sequence.

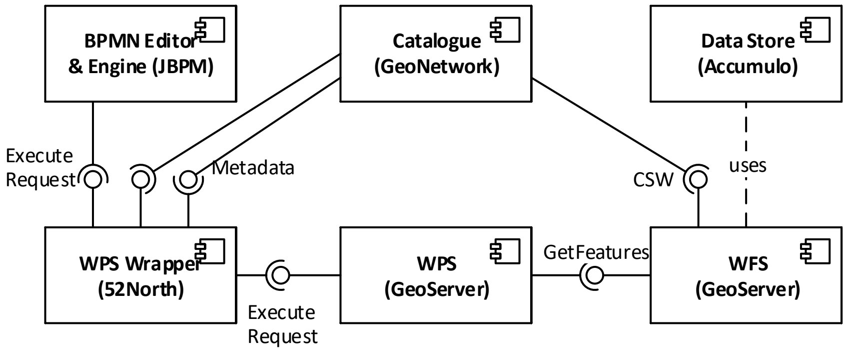

Figure 2. Component diagram of the Web Processing Service profile approach. The implementations used for our testing are shown in brackets.

The implementation of any geoprocessing within this profile will deal first with the metadata record, to request, for example, the data used in the geoprocessing. As a direct consequence, syntactic interoperability is left to the processing step. For example, the different formats available for the dataset therefore have direct access to the metadata which could imply different processing options. Any existing geoprocessing functionality made accessible via WPS can be wrapped into a FMO profiled WPS. This FMO WPS wrapper retrieves the data links from the metadata records and builds the second, non-metadata related WPS request. The workflow architecture using FMO WPS is illustrated in Figure 3 showing the sequence of messages and operations between the different components of the architecture. To highlight what is happening in terms of flow of information (data or metadata) the figure uses an FMO WPS wrapper as described above. Upon execution, the workflow engine begins by constructing an *ExecuteRequest* document comprised of the metadata URL and literal parameters. At this stage no data has been fetched and no 'real' computation has been performed from the workflow engine. Upon execution of the WPS wrapper, the process logic of the wrapper then iterates over each data input and makes a GetRecords request to the catalogue *(Step 1)*. The response of this request is an XML metadata record (ISO 19139). From each record, the URL (*gmd:URL*) is extracted from the distribution information (*MD_DigitalTransferOptions*). If multiple endpoints are listed, we search the list to identify a GML format. A second GetRecords request extracts the metadata record relating to the WPS process (process name and end-point) from the catalogue (*step 2*). The extracted data and process end-point references (*Step 3*) are inserted into a new *ExecuteRequest (Step 4)*, which is executed (*Step 5*) with output(s) specified with the WPS standard *asReference=TRUE* parameter in order to retain the processing result on the server (*Step 6*). The output reference of the WPS is returned to the WPSWrapper which creates a new metadata record containing the output reference within the distribution information tags (*Step 7*). After insertion of the record in the catalogue service, the metadata URL is returned to the workflow client (*Step 8*).

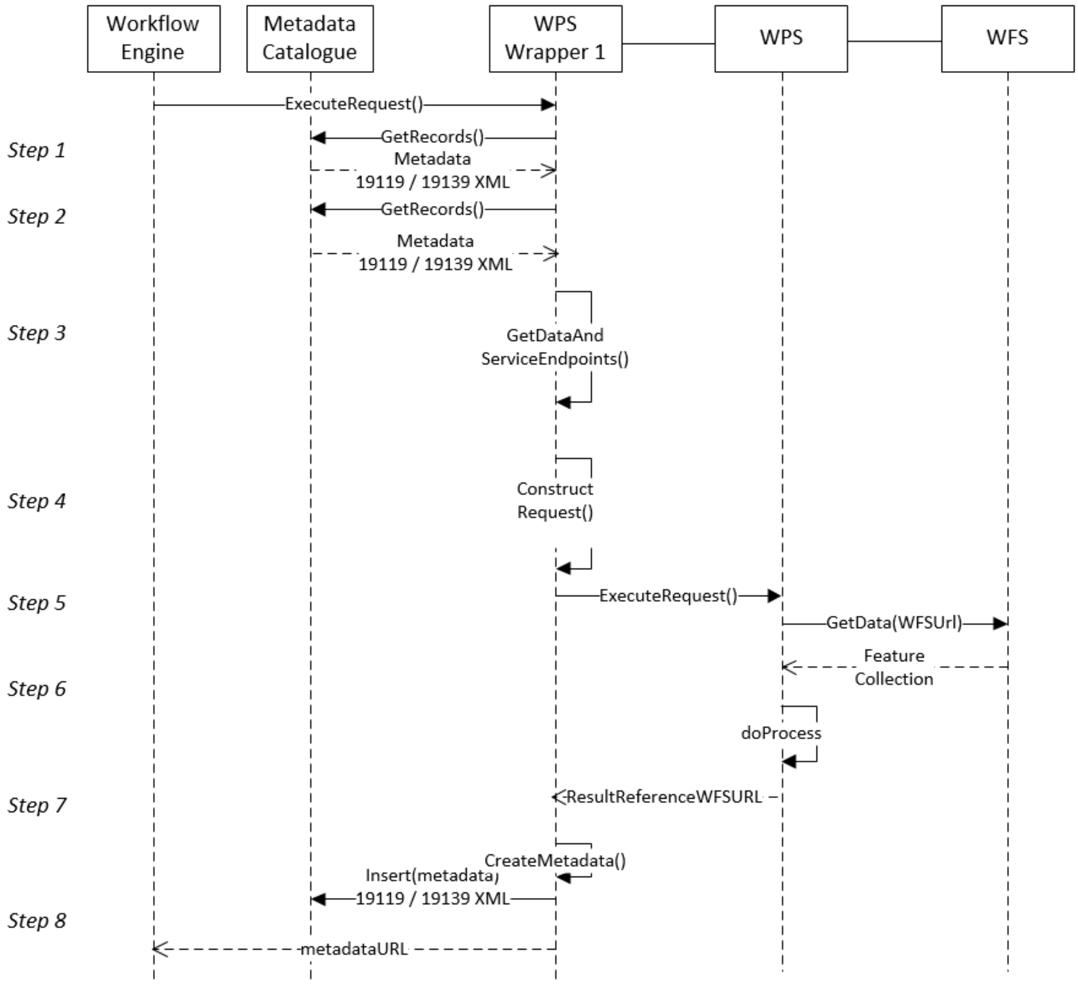

Figure 3. UML sequence diagram of the WPS profiling method.

It is important to notice that without the FMO profiling, the BPMN engine would either get the output as data, thus requiring the presence of the data object in the workflow engine, or as a reference to the result (as part of the WPS standard specification). With FMO, the referencing is masked by being defined just as a string literal that is encoding the URL of the catalogue record. In this architecture, the BPMN engine still has to build the WPS request (as a FMO WPS request) for a task execution. This can be done using a customisation of BPMN for WPS (Meek et al., 2016), or using the BPMN standard specification, where a *servicetask* can be defined as ##WebService with a WSDL association (Sancho-Jiménez, Béjar, Latre, & Muro-Medrano, 2008), if it is available in the workflow engine implementation.

3.2 BPMN profiling

In this configuration, the integration of the metadata catalogue is undertaken at the workflow engine level, rather than as part of a WPS wrapper, as described for the FMO WPS. Figure 4 illustrates the architecture of the system components and Figure 5 shows the execution sequence when using the FMO BPMN profiling.

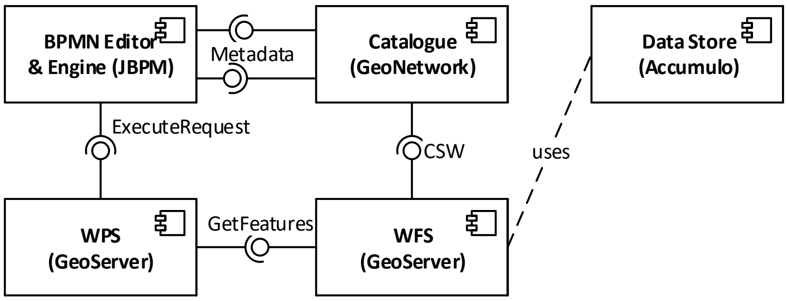

Figure 4. Component diagram of the BPMN profile approach. The implementations used for our testing are shown in brackets.

Here, the integration of the metadata catalogue is undertaken at the workflow engine level, rather than as part of a WPS wrapper, as described for the FMO WPS. However, the same library calls can be utilised from the workflow. Steps 1-8 are equivalent to those described in section 3.1.

The FMO BPMN profiling could be a special case of *servicetask* with a ##WebService parameter defined using WSDL, with the dataset requests made as standard OGC requests (GetData). An alternative is to customise the BPMN editor to support the metadata management. Both solutions keep the BPMN interoperable, but the FMO profiling keeps the details at the metadata level. Also, for semantic support it is desirable to use the BPMN FMO profiling, especially if the workflow editor is capable of providing support based on the knowledge of the metadata. A BPMN editor capable of providing the support can of course also have the capacity to save the BPMN in its basic form, removing the FMO aspects and therefore losing the FMO details in future usage.

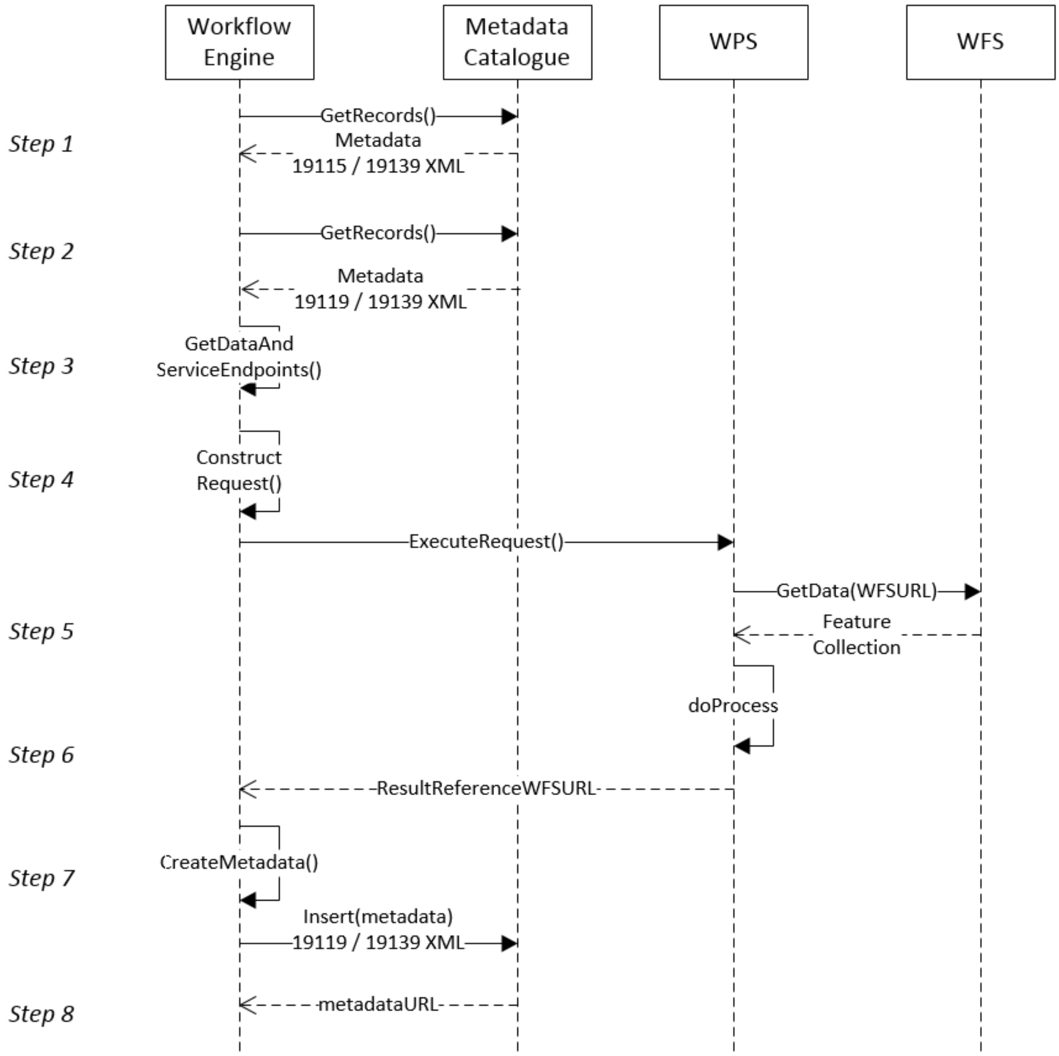

Figure 5. UML sequence diagram of BPMN profiling method.

3.3 Benefits and comparison of FMO solutions

An implementation in the form of a case study big data analysis shows the practical aspects of both FMO profiling methods (see section 4). This is based on a two task workflow to provide a proof of concept of the approaches. The advantage of FMO solutions lies in both the potential semantic support driven by the metadata coupling and that processes do not need to manage datatypes within the workflow and the WPS interface. On one side, matters of semantic and syntactic interoperability (e.g. semantic properties, adequacy, datatype requirements) are dealt at metadata level, with the opportunity to do this within the workflow editor during composition. On the other side, the interoperability requirements are deferred to the last part of orchestration i.e. execution within the WPS. This is obvious for the WPS FMO profiling, but is also the case for the BPMN FMO profiling as it is only at the WPS request that a translation of the metadata to the URL to retrieve the data link is performed and that a registering of the output is also made in the metadata catalogue. In terms of software development, the workflow engine does not need to deal with geospatial data as this is required only within each WPS, i.e. the workflow engine orchestrates tasks which are dealing only with metadata entry points to data records. Spatially-related data such as bounding box and coordinate system definitions are defined as number or string data types.

Both FMO profiling methods exhibit advantages and disadvantages. The advantage of WPS FMO profiling is that the BPMN software for workflow orchestration needs no adaptation beyond making requests and handling responses. In particular, the workflow engine code does not need to be modified to implement the metadata management procedures (which are handled by the wrapper) and instead just passes the metadata URLs. On the other hand, the BPMN profiling solution is more flexible in that it has fewer system components as it does not need an extra WPS wrapper service to be deployed locally with the workflow engine or on another remote system. Thus, it could be easier to configure and maintain, particularly where system administration privileges are restricted or networks are secured. Note that the WPS wrapping mechanism could also be an architecture brokering WPS FMO to a non-FMO BPMN implementation.

3.4 Outlook on FMO solutions for workflows

For both FMO profiling solutions we have described the relative benefits in the previous sections. For either approach, and even for a mixed solution of the two profiles, dealing with the metadata can be advantageous when performed at the workflow level and at WPS level. As mentioned above, each architecture offers flexibility, whilst retaining interoperability, in a similar way to metadata brokering. This focuses mostly on syntactic interoperability, but also offers flexibility at the semantic level through enabling more complex reasoning based knowledge of the manipulated objects when composing or exploring a workflow. This is due to the fact that the BPMN file contains the metadata of all 'objects' used in the workflow: data and (geo) processes.

Therefore composition support can be provided when accessing the metadata, as suggested by Hobona et al (2007), Yue et al (2009) and Al-Areqi et al (2016). For example, attached ontologies could help with harvesting appropriate thematic data with specific characteristics of scale, resolution or quality levels (or other criterion judged to be important) after analysing the metadata related to the task (the process encapsulated in the WPS). Not only does the integration of semantic support into the workflow editor enable the identification of different requirements on compatibility and adequacy, but it can also directly allow different types of workflow execution. A simple example might be verifying the required data format for the process. Similarly, being able to test format availability at the WPS level could lead to a more efficient algorithm (if the WPS allows multiple input formats).

Another example usage is for error propagation analyses. Monte Carlo simulation services can be initiated solely from the information provided by the BPMN file. For example, the file can contain the metadata about the data quality of all datasets used in the workflow (linked via the metadata catalogue service). Therefore it is possible to perform sampling under the given accuracies.

In terms of sharing knowledge, the BPMN file encapsulates all the required information as a single object which can be shared as a scientific model or an application represented by the workflow. In section 4, we demonstrate a simple implementation example of each design we described above to demonstrate FMO workflows.

4 Implementation choices and illustrative example

The workflow implementation adopted here is the jBPM engine and editor environment (github.com/cobweb-eu/workflow-at). The jBPM environment implements version 2 of the BPMN standard. GeoNetwork was chosen for the CSW and GeoServer and 52North were used for WPS processes and FMO WPS wrappers respectively, both of which implement version 1.0 of the OGC WPS standard. Currently, GeoNetwork implements the OGC-CSW 2.0.2 ISO Profile which enables cataloguing of metadata on datasets and services according to ISO19115 and ISO19119 standards (OGC, 2007b). We use the ISO19139 XML schema to encode the metadata records of both the input data (and the process results) objects and processes. The GeoNetwork harvesting module can automatically populate the catalogue with the minimum metadata necessary for the case study datasets. The encoding of the processes as ISO19139 records was achieved through manual definition of the XML, however, the catalogue harvester could be modified to make this process automatic i.e. through invoking GetCapabilities and DescribeProcess operations on the service. In our work, we manually specified the process name (in *MD_DataIdentification*) and WPS endpoint (in *MD_DigitalTransferOptions*) which is the minimum information required in order to run a basic workflow. Additional fields in the ISO19139 schema could be populated as part of an organisation's metadata creation procedures if it is desirable. For example, data quality information might be specified on the inputs (*DQ_DataQuality*), or the lineage field (*LI_Lineage*) might be populated to self-document workflows and enable tracing back through the inputs used.

4.1 Workflow environment

The BPM tool was modified to create and execute WPS requests and manage parsing of the responses to enable process chaining. In jBPM, this is achieved through the introduction of a domain-specific processing task, termed as a *custom work item* or *custom service node*. Although the BPMN standard allows the use of web services through defining a *servicetask* and associating it with a WSDL file, this was not implemented in jBPM. Instead each instance of this custom work item corresponds to a WPS request, with the inputs and outputs of the WPS mapped into a corresponding input and output declaration for the engine. Thus, in this work, each task is defined using MVFLEX Expression Language (MVEL) within the workflow engine and registered with a generic handler for executing a WPS as a BPM item.

Figure 6 illustrates the workflow design environment which is provided via a plugin to the Eclipse Integrated Development Environment. The jBPM package also provides a web-based environment for composition of workflows.

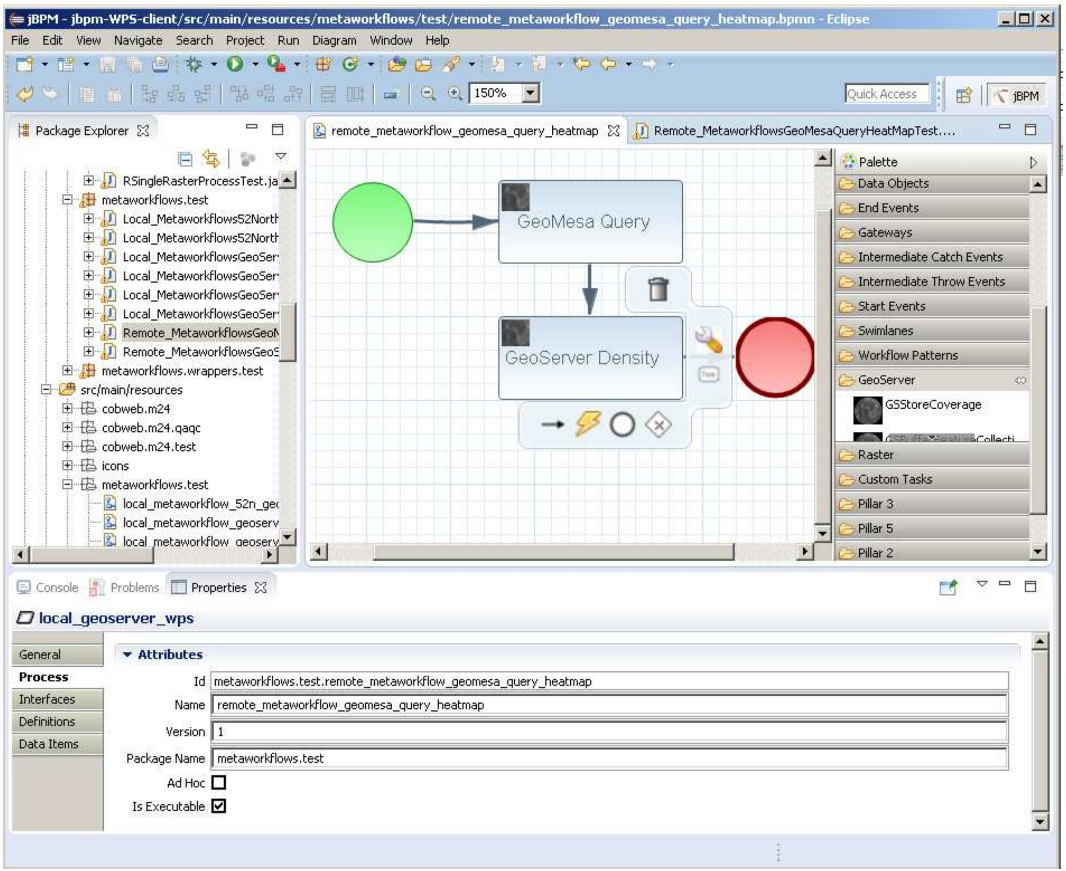

Figure 6. Screenshot of the jBPM workflow editor and development environment.

4.2 Processing and database

In the OGC WPS 1.0 specification, three operations are specified: GetCapabilities, DescribeProcess and Execute. These functions enable a client to identify all the available processes, receive details about the inputs and parameters of specific processes, and invoke running of these operations on the server. Implementation of these functions enables simple integration of data stores and service software. Here GeoServer and its WPS plugin facilitate exposure of a Hadoop back end. Hadoop-based platforms have had recent uptake for undertaking parallelised and large-scale data processing, analysis and storage using clusters of computers and have become a widespread service from cloud computing service providers. While intensive parallel computing is by no means new to geospatial and wider scientific computing applications, extension of the technology is required to support spatial processing. More specifically, a standard Hadoop distribution is not suited to managing geospatial data due its multi-dimensionality. Therefore, a spatial framework is needed to index and handle queries. Partitioning the data for remote sensing applications (Giachetta, 2015) and vector-based query (Whitman, Park, Ambrose, & Hoel, 2014) has been shown and various frameworks have become available as closed and open-source technology stacks such as SpatialHadoop (spatialhadoop.cs.umn.edu). In this work, we adopt the open-source GeoMesa (geomesa.org) project for providing the indexing capability and interface with Hadoop. GeoMesa uses Apache Accumulo (accumulo.apache.org), a column-orientated NoSQL database which enables distributed data storage across a Hadoop cluster (Hughes et al., 2015). GeoMesa also integrates with GeoServer (geoserver.org) enabling implementation and publication of OGC compliant services (including WFS, WCS, WMS and WPS).

4.3 Metadata and metadata catalogue

As a metadata catalogue service, we use GeoNetwork (geonetwork-opensource.org) v3.0.4.0. Currently, GeoNetwork implements the OGC-CSW 2.0.2 ISO Profile which enables cataloguing of metadata on datasets and services according to ISO19115 and ISO19119 standards (OGC, 2007a). We utilise CSW-ISO profile to encode metadata regarding both the datasets and the individual processes that make up a workflow.

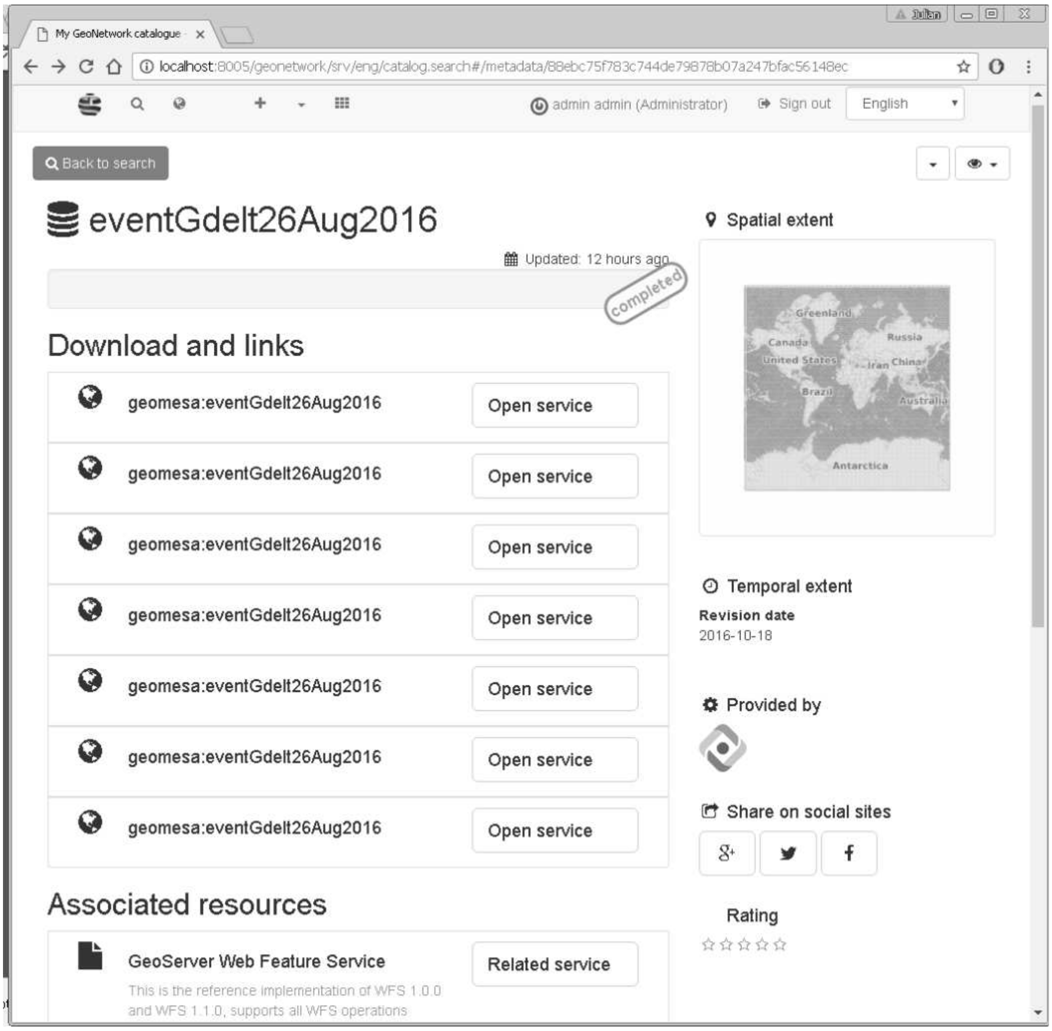

Figure 7. GeoNetwork catalogue application after harvesting GeoServer.

5 Case study scenario

5.1 A big data processing example

Our implementation was tested using a workflow focused on a global event mapping task. As a data source we use the GDELT database which provides a large repository of georeferenced political event data covering 1979 to the present day (Leetaru & Schrodt, 2013). GDELT draws its observations from textual analysis of international news coverage which is automatically processed to create coded observations. The dataset is used for global scale political monitoring and prediction and has, for example, been adopted for forecasting civil unrest (Korkmaz et al., 2015) and monitoring sentiment toward political ideas or events (Bodas-Sagi & Labeaga, 2016).

The reliance of GDELT on automated text analysis means that events may be misclassified, inaccurately geocoded or misrepresentative in other ways and the system has drawn criticism over its validity (Wang, Kennedy, Lazer, & Ramakrishnan, 2016). Therefore, we argue that effective cataloguing of the analysis results and the fact that the complete workflow is documented (via the BPMN definition, which encapsulates the relevant metadata and datasets) makes a relevant scenario for demonstrating the techniques in this paper. Furthermore, the size of GDELT makes query and analysis of the repository challenging. When dealing with such large datasets, network transfer needs to be minimised to ensure timely workflow execution.

The GDELT database uses Conflict and Mediation Event Observation (CAMEO) codes for attributing events. Of particular note is the event code (EventRootCode) assignment which hierarchically classifies the records into "actions" relating to the identified political activity of the event and the actors involved e.g. "Provide Aid" or "Engage in Diplomatic Cooperation". Mapping of these events using density methods provides a way to effectively monitor the spatial distribution of very large datasets and highlight areas for further exploration (Maciejewski et al., 2010). Density based mapping techniques are also relevant to the analysis of sentiment or tone of GDELT event data as investigated by Shook et al (2012).

5.2 Processing workflow details

Figure 8 illustrates the BPMN of the GDELT mapping workflow. The workflow comprises two processes for extracting data and creating the resulting density surface. For this case study, GDELT event data was ingested into Accumulo via the GeoMesa framework as a static loading step undertaken prior to the workflow execution. The data was then exposed as WFS through as GeoServer data store. Limitations in the cluster hardware available meant the input GDELT data was limited to 100,000 features. In practical deployment, this ingestion could easily be automated as part of a regular batch loading or an ongoing streaming job could be implemented. In our implementation test, we configured the workflow Eclipse editor and engine (jBPM) and metadata catalogue (GeoNetwork) on the same local machine. For the FMO WPS testing, the profile wrappers WPS (52North) were also implemented on the local machine. The data and processing service (GeoServer version 2.8.1 and GeoMesa version 1.2.5) was configured on a remote cloud service, as part of a single-node Hadoop (version 2.7.1), Accumulo (version 1.6.5) and Spark (version 1.5.2) cluster.

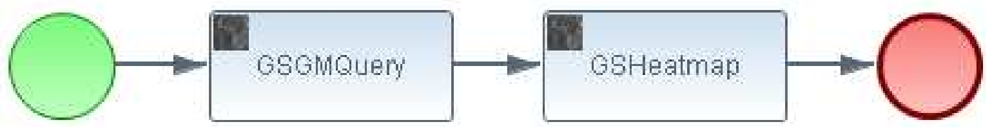

Figure 8. GDELT mapping workflow (BPMN FMO profile) comprising a query WPS (GSGMQuery) and a heat map WPS (GSHeatmap).

The first process of the workflow is a query on GeoMesa which extracts the relevant data according to temporal and event type constraints. Although a similar query could be achieved through specifying a filter in the WFS request (potentially more computationally efficiently), this would require customisation of the workflow editor in order to allow the set the filter parameters in that request. The GSGMQuery process requires two parameters: one for the input metadata record; one for an OGC filter that describes the relevant constraints. For example, once inserted a metadata record may be referred to by its URL in the catalogue e.g. http://localhost:8005/geonetwork/srv/eng/xml.metadata.get?id=46092

The OGC filter defines the relevant event code and data for the GDELT query, see Figure 9. This query expression is provided by the workflow author during composition. Such a query might be taken from a library of pre-set expressions made available to the workflow author.


```xml
<ogc:Filter xmlns:ogc="http://www.opengis.net/ogc">
  <ogc:And>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>EventRootCode</ogc:PropertyName>
      <ogc:Literal>14</ogc:Literal>
    </ogc:PropertyIsEqualTo>
    <ogc:PropertyIsEqualTo>
      <ogc:PropertyName>DATEADDED</ogc:PropertyName>
      <ogc:Literal>20160621</ogc:Literal>
    </ogc:PropertyIsEqualTo>
  </ogc:And>
</ogc:Filter>
```


Figure 9. Example OGC filter used as a parameter the for GSGMQuery process to extracting GDELT events data of type EventRootCode 14 (protest events) on 21st June 2016.

A sample of the metadata that is created and inserted in the catalogue after the process is completed is shown in Figure 10. The metadata link is then passed to the next workflow task which extracts the data from the document for input to the GSHeatmap density map process. The result of this process in turn inserted in the catalogue.

```xml
<gmd:MD_DigitalTransferOptions>
  <gmd:onLine>
    <gmd:CI_OnlineResource>
      <gmd:linkage xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:srv="http://www.isotc211.org/2005/srv">
        <gmd:URL>
          http://localhost:8010/wps/RetrieveResultServlet?id=feb80c8d-ff11-4c95-8622-b6c74a32d5c5out.4209557f-555d-4c5d-b474-489976c1b172
        </gmd:URL>
      </gmd:linkage>
      <gmd:protocol>
        <gco:CharacterString>WWW:DOWNLOAD-1.0-http--download</gco:CharacterString>
      </gmd:protocol>
      <gmd:name xmlns:gmx="http://www.isotc211.org/2005/gmx" xmlns:srv="http://www.isotc211.org/2005/srv">
        <gmx:MimeFileType type=""/>
      </gmd:name>
      <gmd:description>
        <gco:CharacterString/>
      </gmd:description>
    </gmd:CI_OnlineResource>
  </gmd:onLine>
</gmd:MD_DigitalTransferOptions>
```


Figure 10. Extract of the *MD_DigitialTransferOptions* metadata after WPS execution has completed and the metadata has been inserted in the catalogue.

5.3 Results

Figure 11 and Figure 12 show the results of global mapping workflow. For the map covering 21st June 2016 we can identify clear hot spots of activity over United States and South Africa. Interrogation of the data confirms that multiple geocoding of GDELT records are apparent in the result sets.

The ability to orchestrate big data type analyses is a key advantage in adopting the use of meta-objects in workflows. The removal of the need for the workflow engine to understand data types enables passing of URI strings between processes, rather than the object itself. When data volumes are large, or bandwidth is restricted and processes and workflow engine are distributed, then this can improve processing times. Table 1 illustrates the processing time for our two proposed approaches against using a customised BPMN engine for reading and writing geospatial data at the workflow level, as proposed by Meek et al (2016). As can be seen, the FMO WPS nor BPMN profile architecture complete in similar processing times. However, using the non-profiled version of BPMN engine is slower. This is due to the passing of data between the engine and the WPS for each execution of a workflow task.

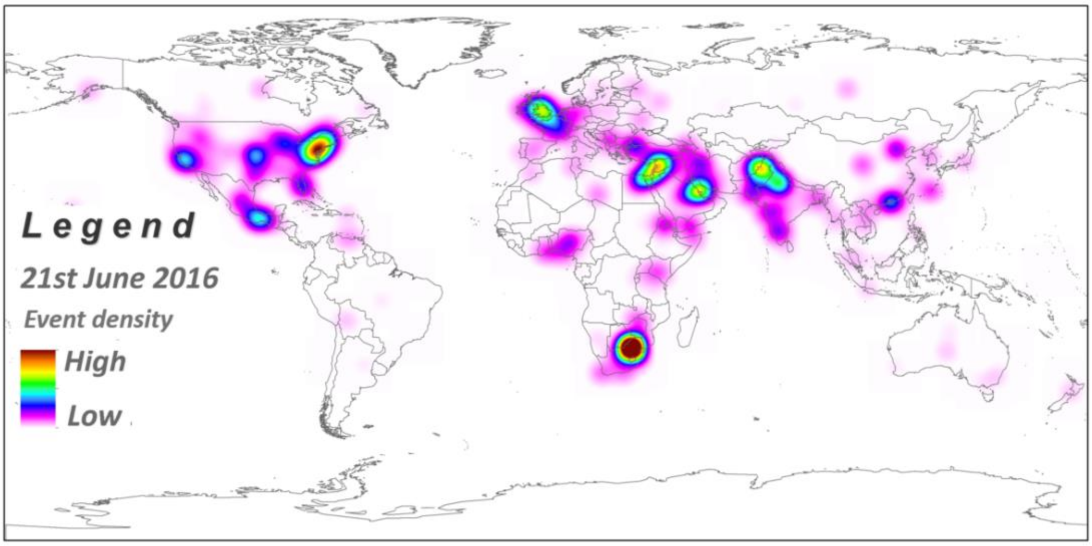

Figure 11. Heat map (kernel density normalised between 0 and 1, 100 pixel radius) of GDELT protest events for 21st June 2016.

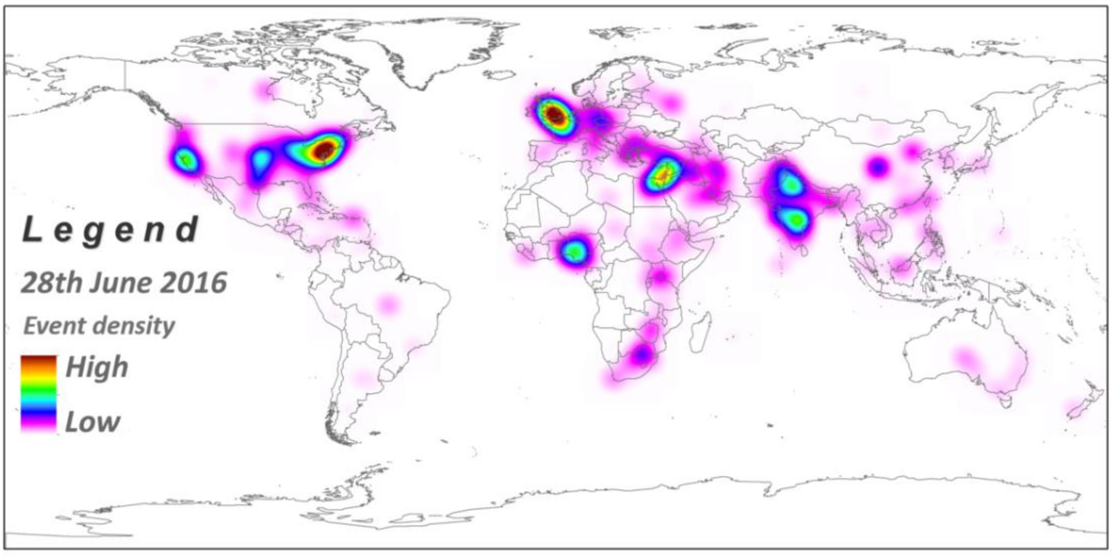

Figure 12. Heat map (kernel density normalised between 0 and 1, 100 pixel radius) of GDELT protest events for 28th June 2016.

Scenario date	***# of features extracted***	***Execution time (seconds)***		
		BPMN FMO profiling	***WPS FMO profiling***	***BPMN (non-profiled)***
21st June 2016	1826	84	93	142
28th June 2016	1352	77	91	139

Table 1. Execution times of the BPMN and WPS Full Meta Object profile approaches and a non-profiled BPMN execution (i.e. WPS invocation using data embedded in the Execute request as proposed by Meek et al (2016)).

6 Conclusion

Orchestrating geoprocessing using an integrated catalogue service aims at facilitating the creation, composition, execution and documentation of scientific geoprocessing workflows. Direct access to metadata of the processes and datasets involved in a workflow eases the syntactic interoperability and if semantic annotations are included in that metadata, could improve the semantic interoperability too. This work developed two novel approaches for coupling a metadata catalogue within a workflow environment and applied them to an analysis workflow comprised of distributed services. Open-source software and standardised interfaces were adopted for this architecture with details of how such approaches can be applied to modern big data analysis platforms.

Several assumptions were made in our approaches with potential disadvantages and areas requiring further research. In one approach, the use of wrappers was presented as a method for avoiding the need to modify existing WPS services. We created these manually for our WPS examples but in practice the use of a broker would be needed to automatically generate the necessary wrapper processes (Boldrini, Papeschi, Santoro, & Nativi, 2015). Furthermore, using the BPMN standard with web service tasks i.e. WSDL with the Full Meta Object profiling methods requires further investigation as the BPM software used did not implement this part of the standard (a method of customisation provided by the platform was adopted instead). This would be important for interoperability when sharing of BPMN files between workflow engines. Lastly, the standardised cataloguing of the complete workflow definition itself was not addressed in this work and would be a valuable area of future work.

The use of a standardised workflow representation together with the potential to integrate processing (both to re-use existing algorithms and exploit new techniques such as big data analysis) is significant, and likely to be of increasing importance as greater numbers of stakeholders in geoprocessing tasks are required to provide input to and share scientific models.

7 References

Al-Areqi, S., Lamprecht, A.-L., & Margaria, T. (2016). Constraints-Driven Automatic Geospatial Service Composition: Workflows for the Analysis of Sea-Level Rise Impacts. In O. Gervasi, B. Murgante, S. Misra, A. M. A. C. Rocha, C. M. Torre, D. Taniar, … S. Wang (Eds.), *Computational Science and Its Applications -- ICCSA 2016: 16th International Conference, Beijing, China, July 4-7, 2016, Proceedings, Part III* (pp. 134–150). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-42111-7_12

Alameh, N. (2003). Chaining Geographic Information Web Services. *IEEE Internet Computing*, *7*(5), 22–29. https://doi.org/10.1109/MIC.2003.1232514

Alonso, G., & Hagen, C. (1997). Geo-Opera: Workflow concepts for spatial processes. In *International Symposium on Spatial Databases* (pp. 238–258).

Bensmann, F., Alcacer-Labrador, D., Ziegenhagen, D., & Roosmann, R. (2014). The RichWPS Environment for Orchestration. *ISPRS International Journal of Geo-Information*, *3*(4), 1334–1351. https://doi.org/10.3390/ijgi3041334

Bielski, C., Gentilini, S., & Pappalardo, M. (2011). Post-disaster image processing for damage analysis using GENESI-DR, WPS and grid computing. *Remote Sensing*, *3*(6), 1234–1250. https://doi.org/10.3390/rs3061234

Bigagli, L., Santoro, M., Mazzetti, P., & Nativi, S. (2015). Architecture of a Process Broker for Interoperable Geospatial Modeling on the Web. *ISPRS International*

Journal of Geo-Information, *4*(2), 647–660. https://doi.org/10.3390/ijgi4020647

Bodas-Sagi, D., & Labeaga, J. (2016). Using GDELT Data to Evaluate the Confidence on the Spanish Government Energy Policy. *International Journal of Interactive Multimedia and Artificial Intelligence*, *3*(6), 38. https://doi.org/10.9781/ijimai.2016.366

Boldrini, E., Papeschi, F., Santoro, M., & Nativi, S. (2015). Enabling interoperability in Geoscience with GI-suite. In *EGU General Assembly Conference Abstracts* (Vol. 17, p. 12199).

Brauner, J., Foerster, T., Schaeffer, B., & Baranski, B. (2009). Towards a research agenda for geoprocessing services. In *12th AGILE International Conference on Geographic Information Science* (Vol. 1, pp. 1–12).

Castronova, A. M., Goodall, J. L., & Elag, M. M. (2013). Models as web services using the Open Geospatial Consortium (OGC) Web Processing Service (WPS) standard. *Environmental Modelling and Software*, *41*, 72–83. https://doi.org/10.1016/j.envsoft.2012.11.010

De Giovanni, R., Williams, A. R., Ernst, V. H., Kulawik, R., Fernandez, F. Q., & Hardisty, A. R. (2016). ENM Components: A new set of web service-based workflow components for ecological niche modelling. *Ecography*, *39*(4), 376–383. https://doi.org/10.1111/ecog.01552

de Jesus, J., Walker, P., Grant, M., & Groom, S. (2012). WPS orchestration using the Taverna workbench: The eScience approach. *Computers and Geosciences*, *47*, 75–86. https://doi.org/10.1016/j.cageo.2011.11.011

Deelman, E., Gannon, D., Shields, M., & Taylor, I. (2009). Workflows and e-Science: An overview of workflow system features and capabilities. *Future Generation Computer Systems*, *25*(5), 528–540. https://doi.org/10.1016/j.future.2008.06.012

Di, L., Shao, Y., & Kang, L. (2013). Implementation of geospatial data provenance in a web service workflow environment with ISO 19115 and ISO 19115-2 lineage model. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(11), 5082–5089. https://doi.org/10.1109/TGRS.2013.2248740

Eberle, J. ., & Strobl, C. . (2012). Web-based geoprocessing and workflow creation for generating and providing remote sensing products. *Geomatica*, *66*(1), 13–26. https://doi.org/10.5623/cig2012-005

Giachetta, R. (2015). A framework for processing large scale geospatial and remote sensing data in MapReduce environment. *Computers and Graphics (Pergamon)*, *49*, 37–46. https://doi.org/10.1016/j.cag.2015.03.003

Henzen, C., Brauner, J., Müller, M., Henzen, D., & Bernard, L. (2015). Geoprocessing appstore. *The 18th AGILE International Conference on Geographic Information Science*.

Hobona, G., Fairbairn, D., Hiden, H., & James, P. (2010). Orchestration of grid-enabled geospatial Web services in geoscientific workflows. *IEEE Transactions on Automation Science and Engineering*, *7*(2), 407–411. https://doi.org/10.1109/TASE.2008.2010626

Hobona, G., Fairbairn, D., & James, P. (2007). Semantically-assisted geospatial workflow design. *Proceedings of the 15th Annual ACM International Symposium on Advances in Geographic Information Systems - GIS '07*, 1. https://doi.org/10.1145/1341012.1341046

Hu, Y., Wu, J., Zhong, H., Lv, Z., & Yu, B. (2010). An approach for integrating geospatial processing services into three-dimensional GIS. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *6318 LNCS*(M4D), 154–161.

https://doi.org/10.1007/978-3-642-16515-3_20

Hughes, J. N., Annex, A., Eichelberger, C. N., Fox, A., Hulbert, A., & Ronquest, M. (2015). GeoMesa: a distributed architecture for spatio-temporal fusion. *Proc. SPIE*. https://doi.org/10.1117/12.2177233

ISO. (2003). ISO 19115:2003 - Geographic Information – Metadata. *Geneva, Switzerland, International Organization forStandardization*.

ISO. (2005). ISO 19119:2005 - Geographic Information – Services. *Geneva, International Standards Organization*.

ISO. (2007). ISO/TS 19139:2007. Geographic information -- Metadata -- XML schema implementation. *Geneva, Switzerland, International Organization forStandardization*.

Kiehle, C., Greve, K., & Heier, C. (2007). Requirements for next generation spatial data infrastructures-standardized web based geoprocessing and web service orchestration. *Transactions in GIS*, *11*(6), 819–834. https://doi.org/10.1111/j.1467-9671.2007.01076.x

Korkmaz, G., Cadena, J., Kuhlman, C. J., Marathe, A., Vullikanti, A., & Ramakrishnan, N. (2015). Combining Heterogeneous Data Sources for Civil Unrest Forecasting. *arXiv Preprint arXiv:1507.05790*, 258–265. https://doi.org/10.1145/2808797.2808847

Leetaru, K., & Schrodt, P. A. (2013). Gdelt: Global data on events, location, and tone, 1979--2012. In *ISA Annual Convention* (Vol. 2).

Maciejewski, R., Rudolph, S., Hafen, R., Abusalah, A., Yakout, M., Ouzzani, M., … Ebert, D. S. (2010). A visual analytics approach to understanding spatiotemporal hotspots. *IEEE Transactions on Visualization and Computer Graphics*, *16*(2), 205–220. https://doi.org/10.1109/TVCG.2009.100

Meek, S., Jackson, M., & Leibovici, D. (2014). A flexible framework for assessing the quality of crowdsourced data. *Proceedings of the AGILE'2014 International Conference on Geographic Information Science*, 3–6. Retrieved from http://repositori.uji.es/xmlui/handle/10234/98927

Meek, S., Jackson, M., & Leibovici, D. (2016). A {BPMN} solution for chaining {OGC} services to quality assure location-based crowdsourced data. *Computers & Geosciences*, *87*. https://doi.org/http://dx.doi.org/10.1016/j.cageo.2015.12.003

Mueller, M., & Pross, B. (2015). OGC WPS 2.0 Interface Standard, 1–133.

Müller, M. (2015). Hierarchical profiling of geoprocessing services. *Computers and Geosciences*, *82*, 68–77. https://doi.org/10.1016/j.cageo.2015.05.017

Müller, M., Bernard, L., & Kadner, D. (2013). Moving code - Sharing geoprocessing logic on the Web. *ISPRS Journal of Photogrammetry and Remote Sensing*, *83*, 193–203. https://doi.org/10.1016/j.isprsjprs.2013.02.011

Nativi, S., Mazzetti, P., & Geller, G. N. (2013). Environmental model access and interoperability: The GEO Model Web initiative. *Environmental Modelling and Software*, *39*, 214–228. https://doi.org/10.1016/j.envsoft.2012.03.007

OGC. (2007a). Open Geospatial Consortium OpenGIS Catalogue Services Specification 2.0.2 - ISO Metadata Application Profile, 7–45.

OGC. (2007b). OpenGIS Catalogue Services Specification, (OGC 07-006r1), 7–45.

OGC. (2007c). OpenGIS Web Processing Service, (OGC 05-007r7).

Oinn, T., Greenwood, M., Addis, M., Alpdemir, M. N., Ferris, J., Glover, K., … Wroe, C. (2006). Taverna: Lessons in creating a workflow environment for the life sciences. *Concurrency Computation Practice and Experience*, *18*(10), 1067–1100. https://doi.org/10.1002/cpe.993

Peltz, C. (2003). Web Services Orchestration and Composition. *Computer*, *36*(10), 46–

52. https://doi.org/10.1109/MC.2003.1236471

Rosser, J., Pourabdollah, A., Brackin, R., Jackson, M., & Leibovici, D. (2016). Full Meta Objects for flexible geoprocessing workflows: profiling WPS or BPMN? In *The 19th AGILE International Conference on Geographic Information Science*.

Sancho-Jiménez, G., Béjar, R., Latre, M. A., & Muro-Medrano, P. R. (2008). A method to derivate SOAP interfaces and WSDL metadata from the OGC web processing service mandatory interfaces. In *International Conference on Conceptual Modeling* (pp. 375–384).

Sheng, Q. Z., Qiao, X., Vasilakos, A. V., Szabo, C., Bourne, S., & Xu, X. (2014). Web services composition: A decade's overview. *Information Sciences*, *280*, 218–238. https://doi.org/10.1016/j.ins.2014.04.054

Shook, E., Leetaru, K., Cao, G., Padmanabhan, A., & Wang, S. (2012). Happy or not: Generating topic-based emotional heatmaps for culturomics using CyberGIS. *2012 IEEE 8th International Conference on E-Science, E-Science 2012*. https://doi.org/10.1109/eScience.2012.6404440

Sun, Z., & Yue, P. (2010). The use of Web 2.0 and geoprocessing services to support geoscientific workflows. *2010 18th International Conference on Geoinformatics, Geoinformatics 2010*, 1–4. https://doi.org/10.1109/GEOINFORMATICS.2010.5567702

Wang, W., Kennedy, R., Lazer, D., & Ramakrishnan, N. (2016). Growing pains for global monitoring of societal events. *Science*, *353*(6307), 1502–1503. https://doi.org/10.1126/science.aaf6758

Whitman, R. T., Park, M. B., Ambrose, S. M., & Hoel, E. G. (2014). Spatial Indexing and Analytics on Hadoop. *Proceedings of the 22Nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 73–82.

https://doi.org/10.1145/2666310.2666387

Wiemann, S. (2016). Formalization and Web-based Implementation of Spatial Data Fusion. *Computers & Geosciences*, *99*(October 2016), 107–115. https://doi.org/10.1016/j.cageo.2016.10.014

Yang, C., Raskin, R., Goodchild, M., & Gahegan, M. (2010). Geospatial Cyberinfrastructure: Past, present and future. *Computers, Environment and Urban Systems*, *34*(4), 264–277. https://doi.org/10.1016/j.compenvurbsys.2010.04.001

Yu, G. (Eugene), Zhao, P., Di, L., Chen, A., Deng, M., & Bai, Y. (2012). BPELPower—A BPEL execution engine for geospatial web services. *Computers & Geosciences*, *47*, 87–101. https://doi.org/10.1016/j.cageo.2011.11.029

Yue, P., Di, L., Yang, W., Yu, G., Zhao, P., & Gong, J. (2009). Semantic web services-based process planning for earth science applications. *International Journal of Geographical Information Science*, *23*(9), 1139–1163. https://doi.org/10.1080/13658810802032680

Yue, P., Gong, J., & Di, L. (2010). Augmenting geospatial data provenance through metadata tracking in geospatial service chaining. *Computers and Geosciences*, *36*(3), 270–281. https://doi.org/10.1016/j.cageo.2009.09.002

Yue, P., Gong, J., Di, L., Yuan, J., Sun, L., Sun, Z., & Wang, Q. (2010). GeoPW: Laying Blocks for the Geospatial Processing Web. *Transactions in GIS*, *14*(6), 755–772. https://doi.org/10.1111/j.1467-9671.2010.01232.x