WindS@UP: The e-Science Platform for WindScanner.eu

Filipe Gomes[1], João Correia Lopes[2], José Laginha Palma[1] and Luís Frölén Ribeiro[3]

[1] FEUP, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
[2] INESC TEC, DEI, FEUP, Universidade do Porto, Rua Dr. Roberto Frias, s/n, 4200-465 Porto, Portugal
[3] IPB, Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal

Abstract.

The WindScanner e-Science platform architecture and the underlying premises are discussed. It is a collaborative platform that will provide a repository for experimental data and metadata. Additional data processing capabilities will be incorporated thus enabling *in-situ* data processing. Every resource in the platform is identified by a Uniform Resource Identifier (URI), enabling an unequivocally identification of the field(s) campaign(s) data sets and metadata associated with the data set or experience. This feature will allow the validation of field experiment results and conclusions as all managed resources will be linked. A centralised node (Hub) will aggregate the contributions of 6 to 8 local nodes from EC countries and will manage the access of 3 types of users: data-curator, data provider and researcher. This architecture was designed to ensure consistent and efficient research data access and preservation, and exploitation of new research opportunities provided by having this "Collaborative Data Infrastructure". The prototype platform—WindS@UP—enables the usage of the platform by humans *via* a Web interface or by machines using an internal API (Application Programming Interface). Future work will improve the vocabulary ("application profile") used to describe the resources managed by the platform.

1. Introduction

This work describes the design of the *WindScanner.eu* e-Science platform and its validation using a prototype—the WindS@UP.

1.1. The WindScanner

One of two main technologies of coherent Doppler lidars are pulsed coherent Doppler or continuous-wave (CW) coherent Doppler scanning lidar. The CW technology is used for the short-range WindScanner and the pulsed coherent Doppler type is used for the long range WindScanner. The whole WindScanner system recurs to three units synced and coordinated to measure the flow field along a certain path in the measuring volume. This is the core technology of the WindScanner infrastructure[1].

The *WindScanner.eu*—The European WindScanner Facility, is an ESFRI project, under the FP7-Infrastructures-2012-1; it is a mobile, distributed, facility with 6 to 8 nodes in European countries. The project aims to be an Open Access distributed research infrastructure promoting the dissemination of results including innovation products and their exploitation [2].

WindScanners are deployed at existing or planned test facilities, covering different climate conditions and terrains. Each *Campaign Site* handles low-level communication with several

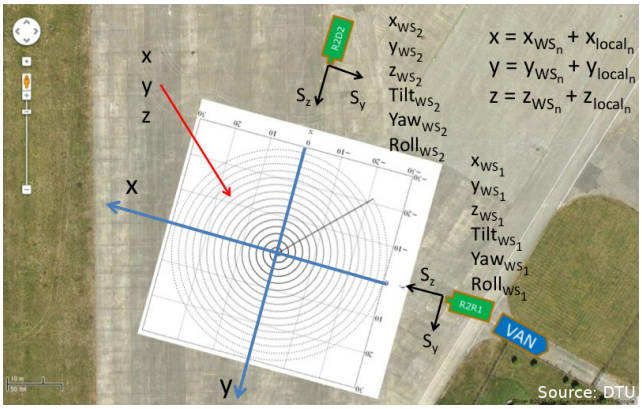

Figure 1. Campaign metadata using two short-range coherent lidars (R1D1 and R2D2) relative to the measuring plane

WindScanner units at the measurement site, Figure 1. One unit, sampling at 400 points produce approximately 6.235 MiB/min (1 Mib=2^{20} bytes) and the motion system (i.e. scanner head, etc.) produces 5.125 MiB/min, given a total for the 3 units of 34 GiB/day ((3*5.125 MiB+6.235 MiB)/min) of data to be managed by the platform.

As detailed later, this collaborative platform will provide a central repository for experimental data (data set with three simultaneous measurements of the wind vector's three components) and metadata (e.g. the experimental setup). In addition, data processing capabilities are to be incorporated thus enabling *in-situ* data processing. Every resource in the platform is identified by a Uniform Resource Identifier (URI), enabling researchers to cite them from outside the platform. Data management, hosting servers, hosting of website, administrative office, training of technicians and researchers operating the WindScanners, and training of users are among some of the competences of the Central Facility.

1.2. Research data management

Many scientific disciplines are increasingly dependent on data, namely the result of a flood of scientific data from a new generation of experiments, simulations, sensors and satellites [3]. Moreover, important scientific research requires a high degree of collaboration and sharing of data. Multiple disciplines come together and research teams are much larger, demanding a shift from working alone with their own data to working in teams with data from multiple sources and, finally, to working in distributed teams with large data sets [4].

Research data management is the management of data produced by an individual or an institution in the course of its research. Research data is very valuable: it is expensive to produce, often useful for entire research communities and supports research findings which can have an impact on society. The emerging "Fourth Paradigm of Science", which comes from the increasingly widespread access to powerful computing capabilities in networked environments, currently allows researchers to produce and manipulate increasingly larger data sets [5]. Thus, research data management is one of the greatest challenges of the information age—a "data deluge" [3] that needs to be addressed to ensure the reproducibility of research findings, allowing the placement of new research questions and increasing the visibility of researchers and institutions [6].

Linked data describes a method of publishing structured data so that it can be interlinked and become more useful [7]. Linked data builds upon standard Web technologies such as HTTP and URIs, but rather than using them to serve Web pages for human readers, it extends them to share information in a way that can be read automatically by computers. This enables data

from different sources to be connected and queried.

Openness is crucial in the research data management environment, because it is a premise for data reuse and also because permissions and licensing management in a data environment would most likely result in a very complex, cumbersome and hard to manage system—perhaps nullifying the advantages of linking resources transparently.

In the following sections, we first discuss related work, analysing similar platforms and survey possible technologies. We then present the different components of the platform, how they relate, and their implementation in the prototype WindS@UP. Finally, we discuss the platform architecture and suggest further work on an elastic infrastructure to achieve scalability.

2. e-Science platforms and projects—an overview

A comparison of existing e-Science platforms and their purpose is presented in Subsection 2.1, while the underlying technologies that may provide support for our platform are also presented and discussed in Subsection 2.2.

2.1. e-Science platforms

There are several projects in e-Science platforms for Data Research Management in European Union, United States and Australia. HUBzero (`http://hubzero.org/`) is an open source software platform for building Web sites that support scientific discovery, learning, and collaboration. Created by researchers at Purdue University to support *nanoHUB.org*, the platform now supports dozens of hubs. HUBzero combines collaborative tools with middleware that provides access to interactive simulation tools. These tools can be designed by any user and range from simulation to data processing. The platform does not currently support Linked Data but some preliminary work exists [8]. Furthermore, Big Data support exists in sites like cceHUB, handling large data sets albeit without automatic handling of received data. While HUBzero provides good support for scientific activities and its simulation tools can be designed with increasing complexity, it does not target data processing.

DataONE (Data Observation Network for Earth) is an infrastructure platform to promote multi-institutional, multinational, and interdisciplinary collaboration. It was developed to support rapid data discovery and access full information life-cycle of biological and environmental data, across geographically distributed data centres and designed to provide researchers with tools that support all elements of the data life cycle—from planning and acquisition through data integration, analysis and visualisation.

The EUDAT project (`http://www.eudat.eu/`) aims to address the challenges of data management and to ensure consistent and efficient research data access and preservation in the scope of the proliferation of data from powerful new scientific instruments, simulations and digitisation of library resources and to exploit new opportunities using the project vision of multi-disciplinary Collaborative Data Infrastructure (CDI). The goal of EUDAT is to build a sustainable cross-disciplinary and cross-national CDI providing a set of shared services to access and preserve research data (such as earth science data). The first set of services provided by the platform are: safe replication, access to data staging for High Performance Computing (HPC), metadata and simple storage.

The e-Science platform to be designed and specified in the *WindScanner.eu* project and later implemented and deployed may follow the same approach as in EUDAT: a "core layer" that may be used by Open Data and any other data and, on top of that, an e-Science and Open Data layer that uses the services the core layer provides.

2.2. Database technologies

A relational database is the predominant choice to store and retrieve data organised in tables.Object-Relational Database Management Systems (ORDBMS) are available in several

scales, varying from embeddable databases to full enterprise-scale databases. ORDBMS have traditionally lagged in supporting core scientific data types, such as N-dimensional arrays. However, these systems are ideal for the management of structured data and metadata as they provide powerful querying mechanisms [9].

The two major and most popular open source ORDBMS are MySQL (`http://www.mysql.com/`) and PostgreSQL (`http://www.postgresql.org/`). Both are mature systems that have a long development history and featuring built-in replication support. This feature is important for the implementation of scalable systems that can accommodate a growing number of campaigns and researchers using the platform. If one of these ORDBMS is used, the database schema may be tuned for better performance, as most of the stored data if immutable (for example a data set representing the collected lidar measurements will not be changed later).

Turning to semi-structured data, NoSQL ("Not only SQL") databases systems trade off consistency, found in traditional relational databases, for higher horizontal scaling and higher availability. These systems generally focus in the retrieval and insertion of data and consist of key-value pair stores. NoSQL databases are suitable for large quantities of data, that do not require the "consistency" properties provided by the relational model, traded with availability and partition tolerance. Apache Cassandra (`http://cassandra.apache.org/`) is a distributed storage system where data is stored in tables consisting of key-indexed multi-dimensional maps.

The requirements of the different scientific domains was determined by a group of science and database experts on issues of: a data model based on multidimensional arrays, not sets of tuples; storage model based on versions and not update in place; open source in order to foster a community of contributors and to ensure that data is never "locked up"—a critical requirement for scientists [10].

With respect to scientific data storage, Array Database Management Systems are tailored specifically for handling arrays, collections of data items on one or more dimensions. Rasdaman (`http://www.rasdaman.com/`) is one such database, that extends standard Relational Database Management Systems with the ability to store and retrieve multi-dimensional arrays.

3. The *WindScanner.eu* e-Science Platform

The platform has a centralised architecture where all data migrates to the Hub node that provides data storage and computing capabilities. The data will be collected at measuring campaign by the windscanners and other wind energy related instrumentation such as, for instance, tower mounted cup or sonic anemometers. The data will be transferred to a local node, local/regional manager of the windscanner facility, who will upload the whole experiment information to the central node (see Figure 2).

This architecture is preferred to a distributed one for two main reasons: data management and data processing. On the one hand, the data published in the platform undergoes a quality assurance process and should be consistently available to any researcher. On the other hand, a centralised infrastructure enables *in-situ* data processing which forgoes the transfer of very large amounts of data over the Internet.

Besides raw and processed data, the platform provides storage for metadata and for other resources—*Research objects*—created by researchers in the course of their work. These encompass a variety of resources, e.g., derived data sets, charts or scientific papers in PDF.

As shown in Figure 2, the experiments' raw data and metadata are transferred from the *Campaign Site* to a *Local Node*. The infrastructure comprises multiple *Local Nodes*, distributed among countries or institutions. Each one collects data from its affiliated *Campaign Sites*, that is, campaigns carried out by a given institution researchers. At this point, a Quality Assurance (QA) process is performed in order to validate the received data. This step may require, depending on the campaign, the processing of raw data to produce "clean data". These data are similar to the raw data—a time series—while some portions may be purged due to

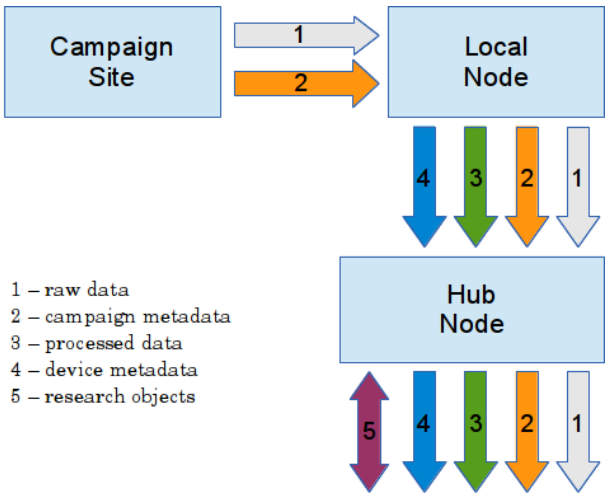

Figure 2. Data flow between nodes

detected errors, or some mathematical operations may be performed. Usually, researchers will use these data rather than the raw data. In addition, further documentation should be added, for instance, detailing the clean data.

The *Hub Node*—the proposed e-Science platform itself—receives these data and metadata, making them available for other researchers. Resorting to these data, they carry on their research, producing *Research Objects*.

During the requirements analysis phase of the project we identified the different user profiles the platform deals with. In general, the platform users are researchers, that want to publish or consult data. Two profiles were defined: *data provider* and *researcher*. The former has data (from carried out experiments) to publish in the platform; the latter search for data and may process data sets found in the platform. A curator will bear the responsibility of validating new data sets so that other researchers may use them with confidence. Users with this third profile—*data curator*—have a crucial role in the platform, intervening after the submission of data. This validation includes checking that data conforms with the defined data quality standards.

The three main components of the e-Science platform are the data storage (Section 3.1), data processing (Section 3.2) and semantic data (Section 3.3).

3.1. Data Storage Platform

As stated previously, scientific research is increasingly dependent on data and requires a high degree of collaboration and sharing of data. Not only are data volumes increasing due to a new generation of experiments and sensors, but also research is shifting to large distributed teams working with large data sets [4, 11].

Two different problems arise given the large amount of research data and the exploitation by large communities: the need to make a researcher's experimental data available to all the others and the need for every researcher to comprehend what the data made available represents. The first problem is fairly straightforward to solve, sufficing a central repository to which every researcher send their data and that everyone consults. The second problem translates to the requirement that data be described in such a manner that a given researcher, which was not involved in its production, understands what they represent and the conditions in which those data were obtained. Further metadata is also preserved, such as authorship, as describer elsewhere.

The authors believe that formats and related standards should be defined prior to the

submission of data sets. In fact, with proper guidelines and formats to which every data provider abides, reasonable data quality should be expected. Moreover, in a given scientific field, it is likely that several sources will provide the same type of data. By using the same formats, a researcher can work with any one of the sources transparently. The definition of relevant formats and data quality standards is outside the scope of the current work but assumed to be present in the platform.

A given data set is defined by means of a data set template, which has two roles: it acts both as the data set *signature* and as a mean to define the content of the data set. The former refers to the data set *mold*, that is, the columns expected in the time series: their name, type and size. In brief, the *signature* is the pattern that the data set must follow. The latter—the definition—consists in a formal description of what each component of the data set represents and how it was obtained. This information may be common to all data sets with a given template, such as error margins in a sensor measurements. On the other hand, the template might define a *parameter* with a precise meaning and a varying value with each data set. These *parameters* allow the presence of common sensor configuration or calibrations, that researchers expect to be available, with specific values for each data set.

Associated with the data set template, and ultimately part of it, the platform enables the storage of documents detailing the experiments (e.g. how experiment was done, what went wrong, the accuracy of instruments, setup and data acquisition). One such is illustrated in Figure 1, in the form of an image. The document presented in the figure describes the position of the several WindScanner sensors. Their relative position to the measuring plane and angles are defined. Each data set template determines exactly what documents must be supplied by the data providers and these documents are subsequently examined in the QA process.

3.2. Data Processing Platform

As a central storage point for experimental data, the platform already proves a great advantage to the work of geographically dispersed researchers in the scientific area. One further advantage, presented in this section, is the possibility of processing data *in-situ*. That is, instead of downloading a data set and processing it locally at researcher's premises, computations for streamline coordinates such as Reynolds Stress $\langle u_i' u_j' \rangle$ or turbulent kinetic energy $\frac{1}{2}\langle u_i' u_i' \rangle$ may be computed locally alongside related data sets.

In-situ data processing presents two major benefits over the traditional procedure. Firstly, data do not need to be moved. In fact, transferring data over the Internet may be very time-consuming for the expected large volumes of data. A researcher is able to select a data set produced by some other entity and comprehend what it represents and the circumstances under which it was obtained.

Secondly, the platform enables sharing of pre-defined or user-defined processing procedures among researchers. On the one hand, this allows for building more advanced procedures iteratively, where different researchers can contribute with new features or optimisations. On the other hand, the availability of well developed and tested procedures proves advantageous to those researchers who are not familiarised with a given data set. The researcher can extract knowledge from the raw data at once, instead of having to analyse the characteristics of data. In many cases, we believe this is preferable because data is only studied superficially. However, one should naturally comprehend fully the specifics of the data when they constitute the basis of the research.

Furthermore, a published procedure can be applied to any data set associated with the supported template. For this reason, a procedure developed targeting a given data set must work—yield results with the same quality expectation—with any other data set with a template in common. Therefore, the procedures should be written once and be applicable to many data sets.

However, as lidars are still growing in the wind industry, some latitude on the database structure must be allowed enabling future research topics on raw data from the lidar such as operations on spectra, operations on line-of-sight speeds or enabling room to accommodate different post-processing techniques for comparison purposes as showed by [12].

In brief, the data processing component of the platform enables researchers to transform data freely. Additionally, the automation of the processing enables the repeatability of computations on different data sets, allowing researchers to focus on their work [13]. The researcher identifies a data set to process, be it raw data or a derived data set. A compatible procedure is then required, that is, a procedure which supports the data set template. More than one procedure may be applied in sequence as long as the output template of the previous and the input template of the next are the same. This way simple workflows are supported.

3.3. Semantic Data Platform

As stated previously, experimental data is both expensive to create and applicable in a wide array of cases. Moreover, documentation of the experiments is imperative for the ensuing use of measured data. Thus, a thorough and patterned description of research objects is crucial for robust data management.

In order to provide long-term preservation, semantic data is stored in the platform in the form of statements comprising a subject, a predicate and an object. The predicates, chosen form an Application vocabulary tailored to the problem domain, are used to describe the resources in the platform. Following the basis of Berners-Lee's Linked Data principles, the resources may be differentiated using the provided URI that acts as a permanent identifier of a resource and allows immediate access to the resource data and metadata. This means that once a data set is used and correctly identified by one researcher (or a machine-to-machine program for that matter), it may be retrieved by any other researcher for future validation once it is possible to cite the URI outside the platform.

The semantic layer of the platform, novel in term of the state of the art, is a powerful tool and allow to aggregate for analysis data that share a common vocabulary with each other. For instance, it can provide all wind data from WindScanner for strong convective cases for all flat terrain campaigns available at WindScanner nodes for flat terrain, or provide offshore shear stress data for stable meteorological conditions are among limitless combinations for fundamental and applied wind energy research. These semantic data translate information found elsewhere in the platform as well as knowledge supplied by researchers themselves when describing their resources.

4. Platform Architecture and Implementation

In this section, we discuss the architecture designed for the platform and the different technologies that were assessed and used. In the end, we show an example of a possible interaction with the WindS@UP e-Science platform prototype.

The platform provides an Application Programming Interface (API) to interact with most of its components (see Figure 3). This enables the development of 3rd party applications that garner the capabilities of the e-Science platform. The API can be divided in three areas targeted at: data upload (number 2 in Figure 3), Business logic common operations (n. 6) and access to the Knowledge base comprising the semantic Web resource descriptions (n. 7). Machine programs have broad access to the platform features through the API (n. 9) and the Web interface we developed uses the API likewise (n. 8).

The storage of scientific data requires a platform that can grow to accommodate an ever increasing amount of data. Also data modification support constitutes a minor problem as, for the most part, data does not suffer updates. Rather, the platform deals with static data—from sensors or derived—that once created, remains preserved. In addition, the storage system must

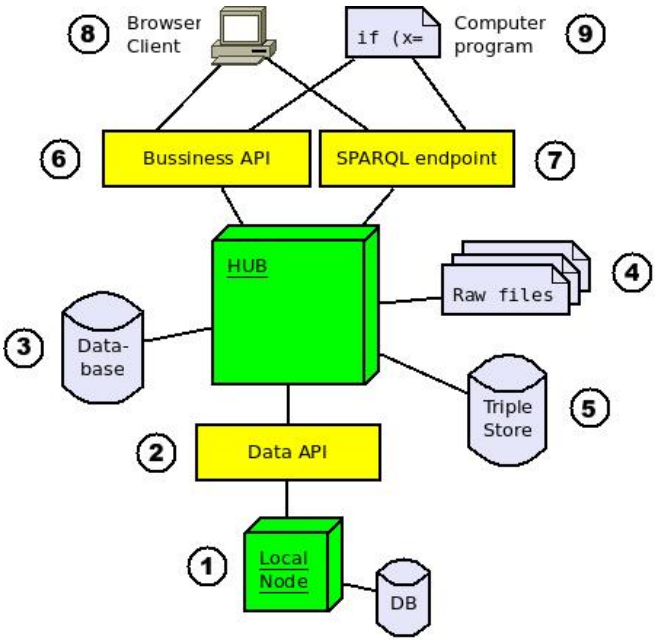

Figure 3. Platform Architecture

be robust so that queries in very large databases run with low overhead. In Section 2, we identified specialised storage systems targeting scientific data, which should fit our needs. Still, we resolved to use PostgreSQL database management system (n. 3) to store operational data: data sets and their metadata. We consider that the alternatives are not yet stable or documented adequately and that we would not be able to take advantage of most features offered. On the other hand, PostgreSQL is a proved, well-known system with a reliable performance and improving scalability.

The triple store (n. 5) for linked data is implemented with Apache Jena (`http://jena.apache.org/`), a Java framework for building linked data applications. It offers an API that handles query processing and abstracts the persistence of the semantic web data. The storage is backed by a relational database, PostgreSQL in the case of our implementation. Jena enables the platform to process SPARQL queries and provides a SPARQL endpoint (n. 7). In the prototype, standard queries were created for what would be recurrent requests. A SPARQL endpoint, is a SPARQL protocol service that enables users—human or machine—to run ad-hoc queries written in the SPARQL language inside the knowledge databases. Moreover, results are typically returned in one or more machine-processable formats. Additional means of access, such as an endpoint for querying semantic Web data directly, enables users to use the access method that best fits their needs [14] and integrates our database in the Linked Open Data Cloud [15].

We should note for future work, however, that this solution using Apache Jena may not scale easily for a very large collection of linked objects. In fact, user queries can demand a relatively large computational power for large semantic databases. Nevertheless, studies have been made to address the scalability of SPARQL queries [16].

Aside from data sets, the platform stores miscellaneous binary large object (BLOBs): raw data, metadata documents, and research objects. These objects, like the data sets, are not modified after creation. Systems such as storage area network (SAN) or network-attached storage (NAS) offer the necessary features and are adequate for our use case (n. 4).

As for data processing, the platform launches background jobs that handle data fetching, processing and subsequent storage. Each processing job has an input and output data set and

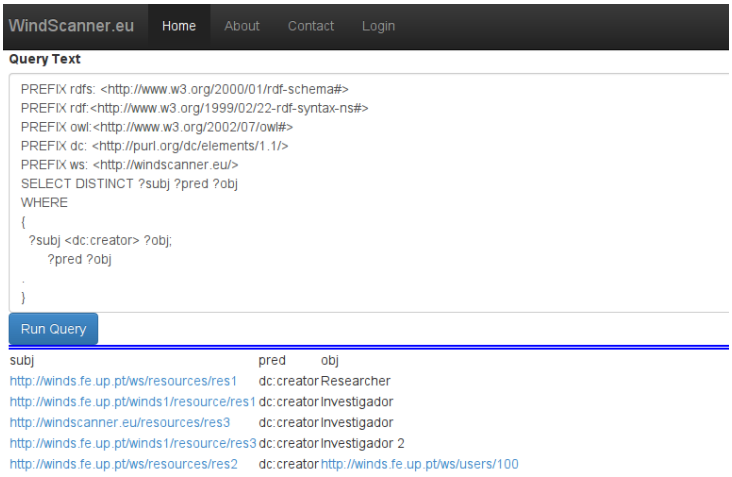

Figure 4. Querying the platform

a processing routine and comply to the defined templates. A support for GNU Octave (`http://www.gnu.org/software/octave/`), a language that is mostly compatible with MATLAB, was also implemented. These jobs just need access to the data repository and may be distributed across multiple machines.

Figure 4 illustrates a possible query a user or machine may pose to the platform API. The text is written is SPARQL and the results shown have resources with links the user may follow to obtain more information from the platform.

5. Conclusion and Future Work

This paper has described the WindScanner e-Science platform prototype (WindS@UP) and its context in the current research problems related with data set (mainly time series) processing and preservation. The proposed research infrastructure with Campaign sites to collect WindScanner data, Local nodes that will manage measurement campaigns and apply data quality procedures to data and a Hub to support the exchange of data and metadata among researchers was detailed.

Researchers' needs are provided by an e-Science platform where they may collect data sets with campaign measurements, run their computations against existing data to produce new data sets, upload other research objects (such as research papers in PDF), describe any research object using a controlled vocabulary (metadata), make comments on any research object and upload procedures to be run against data sets.

The procedures for data quality will be prior to the upload of data (and metadata) to the platform to certify the data to an agreed format before storing it for long term persistence. Apart from the computational needs, the e-Science platform, which stores and provides access to researchers to data from multiple campaigns from several sites, provides elastic storage to accommodate the growing amount of data originated from the equipment and also the data generated by the researches themselves. We facilitate the access to the platform to upload of the various resources (data sets, research objects, etc.) through a set of Web Services, with a documented API, that enforces the verification of access privileges and interoperable data formats.

In the context of the WindScanner.eu project, other tasks will further discuss and complete the issues introduced in this work, namely: devise an Institution responsible to manage the e-Science platform, define user security rules, define upload formats, define I/O formats, define quality control procedures and audits, and setup the platform internal databases. Nevertheless,

it should be noted that the features identified and described in this work are not final and further discussions with the project stakeholders are needed.

For future work one should note that in order to accomplish its purpose, the e-science platform must scale to accommodate the ever-growing experimental and derived data and the users' computing needs. An internal API to isolate the platform from storage and computational needs provides by a cloud layer has been designed and it being is prototyped at the moment. By introducing the internal API in the WindS@UP, we believe it will be easy to provide the scalability needs for the platform in terms of storage and processing power to accommodate big data or a fast growth of the number of users.

Acknowledgements
The *WindScanner.eu*—The European WindScanner Facility—is an ESFRI project (N: 312372) under the FP7-Infrastructures-2012-1. The authors are grateful to all colleagues in WP5 for the fruitful discussions, namely Dimitri Foussekis (CRES), Doron Callies (IWES Fraunhofer), Hans Verhoef (ECN), Harald Svendsen (Sintef), Jan Willem Wagenaar (ECN), Javier Sanz Rodrigo (CENER), Martin Bitter (Forwind), Mikael Sjöholm (DTU), Steen Arne Sørensen (DTU) and Teresa Simões (LNEG).

References

[1] Mikkelsen T, Siggaard Knudsen S, Sjöholm M, Angelou N and Tegtmeier A 2012 *International Conference on Wind Energy: Materials, Engineering, and Policies (Hyderabad, India, Nov 22-23)*

[2] WindScannereu 2012 The European WindScanner Facility: Annex I—Description of Work Tech. rep. Seventh Framework Programme

[3] Hey T and Trefethen A 2003 *Grid Computing: Making The Global Infrastructure a Reality* January 2003 ed Berman F, Fox G and Hey J A (Wiley) pp 809–824

[4] Atkinson M and Roure D D 2009 Data-Intensive Research: making best use of research data (Draft 1) Tech. Rep. December Edinburgh & Southampton

[5] Hey T, Tansley S and Tolle K 2009 *The Fourth Paradigm: Data-Intensive Scientific Discovery* ed Hey T, Tansley S and Tolle K (Microsoft Research, Redmond)

[6] Borgman C L 2011 *Journal of the American Society for Information Science and Technology* 1–40

[7] Bizer C, Heath T and Berners-Lee T 2009 *International Journal on Semantic Web and Information Systems* **5** 1–22

[8] Witt M and Yu Y 2012 *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries* (ACM Press) pp 149–152

[9] Gray J, Liu D T, Nieto-Santisteban M and Szalay A S 2005 Scientific Data Management in the Coming Decade Tech. Rep. Microsoft Technical Report MSR-TR-2005-10 Microsoft

[10] Stonebraker M, Becla J, Dewitt D, Lim K, Maier D, Ratzesberger O and Zdonik S 2009 *Fourth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, Jan 4-7, 2009*

[11] Fox P and Hendler J 2009 *The Fourth Paradigm: Data-Intensive Scientific Discovery* ed Hey T, Tansley S and Tolle K (Microsoft Research, Redmond) pp 147–152

[12] Wagner R, Sathe A, Courtney M, Clifton A, Pedersen T, Mann J and Ribeiro L M F 2014 *17th International Symposium for the Advancement of Boundary-Layer Remote Sensing (Aotearoa, New Zealand, Jan 28-31)*

[13] De Roure D and Goble C 2009 *Software, IEEE* **26** 88–95

[14] Heath T and Bizer C 2011 *Linked Data—Evolving the Web into a Global Data Space* (Morgan & Claypool Publishers)

[15] Bauer F and Kaltenbock M 2012 *Linked Open Data: The Essentials—A Quick Start Guide for Decision Makers* (The Semantic Web Company)

[16] Groppe J and Groppe S 2011 *Proceedings of the 2011 ACM Symposium on Applied Computing* (New York, NY, USA: ACM) pp 1681–1686