The Success of European Projects using New Information and Communication Technologies

Colmar, France
July, 2015

Sofiane Hamrioui (Ed.)

Sponsored and Organized by INSTICC
Published by SCITEPRESS

Edited by Sofiane Hamrioui

Printed in Portugal
ISBN: 978-989-758-176-2
Depósito Legal: 404179/16

Foreword

This book contains the revised and extended versions of papers describing a number of European projects that were presented at the **European Project Space (EPS)** event organized in Colmar, July 2015, associated with the set of conferences ICETE (12th International Joint Conference on e-Business and Telecommunications), ICSOFT (10th International Joint Conference on Software Technologies), SIMULTECH (5th International Conference on Simulation and Modeling Methodologies, Technologies and Applications) and DATA (4th International Conference on Data Management Technologies and Applications).

All these events were sponsored by the Institute for Systems and Technologies of Information, Control and Communication (INSTICC) in collaboration with several international associations and other scientific partners.

The objective of the EPS is to provide a platform and meeting point for researchers and practitioners to discuss and exchange experiences on European Research and Development projects which are funded and developed in the European research area or in collaboration with European partners, thus providing insights into cutting edge research work originating from Academia or Industry, either in Europe or elsewhere.

We aim to establish the EPS as a regular event to create opportunities for knowledge and technology sharing, and establish the basis for future collaboration networks involving current project partners and interested conference delegates.

This event included a panel discussion with representatives from the European Community, namely Dr. Jean-Jacques Bernardini, (Alsace Innovation, France), Dr. Anthony C. Boucouvalas, University of

Peloponnese, Greece and Dr. Ray Walshe (EIURA, Ireland). The EPS technical program included, in addition to an opening panel, the presentation of ten projects which, after the event, have been invited to publish a short report in this EPS book.

We would like to thank the project representatives that decided to take their time and effort to respond to our invitation, whose reports correspond to the chapters of this book.

We would like to express our thanks to the EPS project representatives, who presented their projects in Berlin and took the time to write the chapters of this book, and whose quality work is the essence of the EPS event and of this publication.

Sofiane Hamrioui

Editor

Organization

Panel Chair

Sofiane Hamrioui, USTHB University and UMMTO Univesity, Algeria and UHA University, France

Panel Participants

Jean-Jacques Bernardini, Alsace Innovation, France
Anthony C. Boucouvalas, University of Peloponnese, Greece
Ray Walshe, EIURA, Ireland

Presented Projects

Acronym: H2R
Presenter: Vittorio Lippi, Uniklinik Freiburg, Germany

Acronym: LinDA
Presenter: Anastasios Zafeiropoulos, Ubitech Ltd., Greece

Acronym: ARCADIA
Presenter: Anastasios Zafeiropoulos, Ubitech Ltd., Greece

Acronym: Arrowhead
Presenter: Thibaut Le Guilly, Aalborg University, Denmark

Name: Advanced Museum Services
Presenter: Simone Porru, University of Cagliari, Italy

Acronym: PERICLES
Presenter: Jean-Yves Vion-Dury, Xerox Research Centre Europe, France

Acronym: AMSUN
Presenter: Jens Möckel, Universität der Künste Berlin, Germany

Acronym: DataPipe
Presenter: Florent Bourgeois, Actimage GmbH. Université de Haute-Alsace (UHA), France

Name: Document Management
Presenter: Filippo Eros Pani, University of Cagliari, Italy

Name: Facilitating Industry University Collaboration and Engagement
Presenter: Ray Walshe, EIURA, Ireland

Table of Contents

Papers

Papers

Information Driven Cyber Security Management through LinDA

Panagiotis Gouvas[1], Eleni Fotopoulou[1], Spyros Mouzakitis[2] and Anastasios Zafeiropoulos[1]

[1]UBITECH Ltd., 15231 Chalandri, Athens, Greece
[2]Decision and Support Systems Laboratory, National Technical University of Athens, 15780 Zografou, Athens, Greece
{pgouvas, efotopoulou, azafeiropoulos}@ubitech.eu
smouzakitis@epu.ntua.gr

Abstract. The continuous evolution and adaptation of cybercrime technologies along with their impact on a set of services in various business areas make necessary the design and development of novel methodologies and tools for protecting public organizations and businesses' infrastructure as well as end users. Such methodologies and tools have to exploit the massive power that can be provided through the available data collected in organizations' perimeter, networking and infrastructure level as well as data collected via endpoint devices. However, the available data, in order to be easily exploitable have to be represented in standardized formats or be easily interlinked and being analyzed. Towards this direction, linked data technologies can be used towards the appropriate interconnection of available entities/concepts among different cyber security models. In the current manuscript, we describe an approach for supporting information driven cyber security management through exploitation of linked data analytics technologies, as they are developed within the framework of the LinDA FP7 project. A cyber security data model is designed for cyber-attacks data representation, while a set of insights are produced upon data analysis over data collected in a small enterprise environment.

1 Introduction

Over the last years, the evolution of the Internet along with the emergence and adoption of novel ICT technologies and the development of a portfolio of online services have led to the appearance of a wide set of online threats, risks and vulnerabilities that are handled mostly per case upon their appearance. The arisen security issues impact significantly the daily operation of enterprises and public sector organizations as well as the overall economic growth indicators.

Taking into account the continuous evolution and adaptation of cybercrime technologies as well as the huge number of people that they affect, advanced methodologies and operational tools have to be developed for protecting businesses' and public organizations' infrastructure and citizens. Threats have to be faced at their creation while the overall complexity in the threat management and remediation process has to be minimized. The emergence of cyber security management methodologies and tools

has also to be combined with an increased level of collaboration and exchange of information among enterprises and public organizations, targeting at the increase of their awareness with regards to the design and implementation of cyber security solutions as well as the facilitation of the specification of effective policies for handling cyber threats.

Towards this direction, a set of challenges is identified. These challenges regard (i) the efficient processing of the available cybersecurity-oriented information from internal and external sources within an enterprise/organizational environment (e.g. raw data with regards to incidents, vulnerabilities, weaknesses etc.), (ii) the extraction of advanced knowledge upon the available cyber security information based on the application of a set of knowledge-extraction and management algorithms, (iii) the application of effective and efficient mechanisms for cyber-security management by making use of the available information flows and taking in parallel into account constraints with regards to peculiarities imposed by the provided services, (iv) the design of user-friendly tools for information driven cyber security management that facilitate timely and efficient response to incidents without the need for actions from cybersecurity specialised personnel and (v) the promotion of unified open cyber security data publication schemes along with interoperability mechanisms to be used/consumed by enterprises and public organizations.

Processing of the available cybersecurity-oriented information has to be realised from internal and external sources within an enterprise/organizational environment – through cyber security monitoring tools- as well as over raw data or data available in heterogeneous formats. Processing of such information can lead to efficient decision making in real time as well as a posteriori with regards to the implementation of cyber-security solutions. In order to deploy efficient information processing schemes, advanced techniques have to be applied for information representation as well as concepts interconnection processes. Efficient representation necessitates the existence/usage/extension of commonly used cyber security meta-models, as well as the application of mapping mechanisms for transformation of the available data to formats that can be easily and commonly processed. Such a mapping can be realised through the development of an ontology (or group of ontologies) of the cyber security domain, expressed in a specific language (e.g. OWL language), that will enable data integration across disparate data sources. Formally defined semantics will then make it possible to execute precise searches and complex queries and support semantically alignment processes among datasets represented by different models.

Towards this direction, linked and open data technologies can be exploited. The term linked data refers to a set of best practices for publishing and interlinking structured data on the Web. By following these practices, data from diverse sources can use the same standard format that allows them to be combined and integrated. Linked data specify that all data will be represented based on the Resource Description Framework (RDF) specifications. Conceptual description of data is realized based on specific vocabularies (and thus semantics) accessible over HTTP, allowing the user to interpret data from multiple vocabularies and query them in a uniform manner. By adopting linked data principles, a set of advantages are provided towards the production of advanced analytics and insights. Combination of data from multiple and in many cases distributed sources, as well as from publicly available data (e.g. open governmental data) or privately owned data maintained by enterprises, can help businesses enhanc-

ing their experience of managing and processing of data, in ways not available before. Actually, linked data provide the capacity for establishing association links among concepts in different datasets, producing high-quality interlinked versions of semantically linked web datasets and promoting their use in new cross-domain applications by developers across the globe. Such interlinked datasets constitute valuable input for the initiation of an analytics extraction process and can lead to the realization of analysis that was not envisaged in the past.

In the cyber security domain, linked data can be used towards the appropriate interconnection of available entities/concepts among different cyber security models. Linked data analysis provides cyber experts and incident responders a way to quickly identify the important assets, actors, and events relevant to their organization, accentuating the natural connections between them and providing contextual perspective. With this added context, it becomes much easier to see abnormal activity and assess the blast radius of an attack [1]. However, the power of linked data can be fully exploited, given the existence of significant amount of data, made available by public organizations and enterprises. Open data publication and consumption schemes have to be adopted and widely used for the aggregation of cyber security associated data in open repositories. Over such repositories, queries on the available open data or interlinking of data for advanced queries can be applied. The wide adoption of open data technologies can facilitate the appropriate dissemination of information with regards to new threats and vulnerabilities, the realisation of advanced analysis taking into account available data from other sources as well as the shaping of communities of practice and the engagement of "non-experts" in the cyber security domain.

Extraction of knowledge and management of the available information upon the mapped/interlinked data can be realised through the application of novel analysis techniques as well as the development of user-friendly analytics and visualisation tools. Novel analytic and visualisation approaches have to be introduced and provided to end users through user-friendly tools. Analysis has not only to focus on extraction of conclusions and results based on experiences from previous threats, attacks and risks. A set of analytics for identification of malicious behaviours, anomaly detection, identification of epidemiological incidents etc. has to be supported even for decisions that have to be made in real time. This is not to say that preventive measures are useless, but instead that organizations must arm themselves with proficient detection and response practices for readiness in the inevitable event that prevention fails [1].

Going one step further, such tools have to support functionalities for the extraction of linked data analytics [2], given that analytics are in most cases related with the processing of data coming from various data sources that include structured and unstructured data. In order to get insight through the analysis results, appropriate input has to be provided that in many cases has to combine data from diverse data sources (e.g. data derived from endpoints in different geographical areas). Thus, there is inherent a need for applying novel techniques in order to harvest complex and heterogeneous datasets, turn them into insights and make decisions.

Taking into account the afore-mentioned challenges and enabling technologies for overcoming part of them, it could be claimed that there is open space for the design, development and validation of novel information driver cyber security management solutions that can unleash the potential of the processing of huge amount of the available information. In the current manuscript, such an approach is presented based on

the realisation of linked data analysis through the workbench that is developed within the framework of the FP7 project LinDA (http://linda-project.eu/). A cyber security data model is designed for cyber-attacks data representation, while through the usage of the LinDA workbench, data transformation to RDF format and data analysis is supported. Available data stem from data collected through monitoring of cyber-attacks in a small enterprise environment. The produced insights of the analysis, along with the definition of the cyber security data model, constitute outcomes that can be the basis for further extensions and more advanced analyses in the future.

In more detail, the structure of the paper is as follows: in section two, the LinDA project including its main objectives, the LinDA workflow and the developed LinDA workbench is described; in section three a pilot application scenario in the cyber security domain is presented along with the analysis results and the produced insights, while section four concludes the paper by referring to the exploitable outcomes of the presented work and plans for future work.

2 The LinDA Project

LinDA [3] aims to assist SMEs and data providers in renovating public sector information, analysing and interlinking with enterprise data by developing an integrated, cross-platform, extensible software infrastructure, titled as the "LinDA workbench" [4] that handles the end-to-end the transition of a data-powered enterprise to a linked data-powered organisation. The LinDA workbench allows for the transformation of various formats of data into arbitrary RDF graphs, the construction of linked data queries through user friendly interfaces addressed to SMEs, included intuitive visualisation methods and charts, while it is in a position to perform further statistical analysis to the queries results based on the R framework [5]. The LinDA workbench can be used either as a service, or be deployed and operated as a standalone solution, while users are in a position to make use of its sub-modules in a distinct manner also.

The overall realisation of the LinDA project has been achieved through the realisation of the following objectives:

- enhance the ability of data providers, especially public organisations to provide re-usable, machine-processable linked data.
- provide out-of-the-box software components and analytic tools for SMEs that offer the opportunity to combine and link existing public sector information with privately-owned data in the most resourceful and cost-effective manner.
- deliver an ecosystem of linked data publication and consumption applications that can be bound together in dynamic and unforeseen ways.
- demonstrate the feasibility and impact of the LinDA approach in the European SMEs Sector, over a set of pilot applications.
- achieve international recognition and spread excellence for the research undertaken during the LinDA implementation towards enterprises, scientific communities, data providers and end-users. Diffuse and communicate readily-exploitable project results, of a pro-normative nature. Contribute to standardisation and education.

2.1 The LinDA Workbench

The LinDA workbench [4] concerns an open-source package of linked data tools for enterprises to easily publish data in the linked data format, interlink them with other data, analyze them and create visualizations. The main components of the LinDA workbench (Figure 1) are the following:

- The LinDA Transformation Engine, a data transformation solution that provides a simplified workflow for renovating and converting a set of common data containers, structures and formats into arbitrary RDF graphs. The Transformation Engine can be used to develop custom solutions for SMEs and public sector organisations or be integrated into existing open data applications, in order to support the automated conversion of data into linked data. The overall platform allows the export of arbitrary RDF graphs as tabular data, supporting SMEs to store the final results of data linking into relational databases or process further with spreadsheet and data analysis software.

- The LinDA Vocabulary Repository, a repository for accessing and sharing linked data that can be linked to the Linked Open Data (LOD) cloud. The system allows SMEs to reference and enrich metadata shared by well-established vocabulary catalogues (e.g. LOV, prefix.cc, LODStats), thus contributing to easy and efficient mapping of existing data structures to the RDF format as well as to increasing the semantic interoperability of the SMEs datasets.

- The LinDA Query Designer and the Query Builder tools that enable non-experts to formulate a SPARQL query and explore open datasets in an innovative and easy way, to use graphical methods to interactively build a simple or complex query over multiple data sources and view the results in a SPARQL editor. The Query Designer follows the paradigm and quality of SQL Query designers of popular relational database management systems where, with simple drag'n'drop functionality, users can perform complex SPARQL queries, while the Query Builder offers similar functionality through a wizard-like guided list procedure.

- The LinDA Visualization engine that can help enterprise users gain insight from the linked data that the company generates. With this engine users can visualize data in linked data format taking into advantage their semantics. The LinDA visualization provides a largely automatic visualization workflow that enables SMEs to visualize data in different formats and modalities. In order to achieve this, a generic web application is being developed based on state-of-the-art linked data approaches to allow for visualizing different categories of data, e.g. statistical, geographical, temporal, arbitrary data, and a largely automatic visualization workflow for matching and binding data to visualizations.

- The LinDA Analytics and Data Mining component [2] supports the realization of analysis based on the consumption and production of linked data. A library of basic and robust data analytics functionality is provided through the support of a set of algorithms, enabling organizations and enterprises to utilize and share analytic methods on linked data for the discovery and communication of meaningful

new patterns that were unattainable or hidden in the previous isolated data structures. The analytics and data mining component is based on an extensible and modular architecture that facilitates the integration of algorithms on a per request basis. The development of the component is based on open-source software while integration of algorithms is based on open-source analytics projects (mainly, the R statistics project).

- An ecosystem of linked data consumption applications, which can be bound togeth+er in a dynamic manner, leading to new, unpredicted insights. The consumption applications regard a set of applications that are developed aiming to provide to end users (including pilot users) functionalities that are not provided through the LinDA Workbench. The objective is to facilitate, through specific applications, the daily business processes of the SMEs based on the redesigned workflows that take into account the usage of the LinDA tools. As such, the consumption applications can be considered as small tailor made solutions, easy-to-implement and of low-cost, serving the specific needs of the LinDA end-users that occurs while interacting with LinDA workbench. In this way, the LinDA workbench can be considered as a complete, end-to-end solution for the incorporation of the linked data concepts within the SMEs.

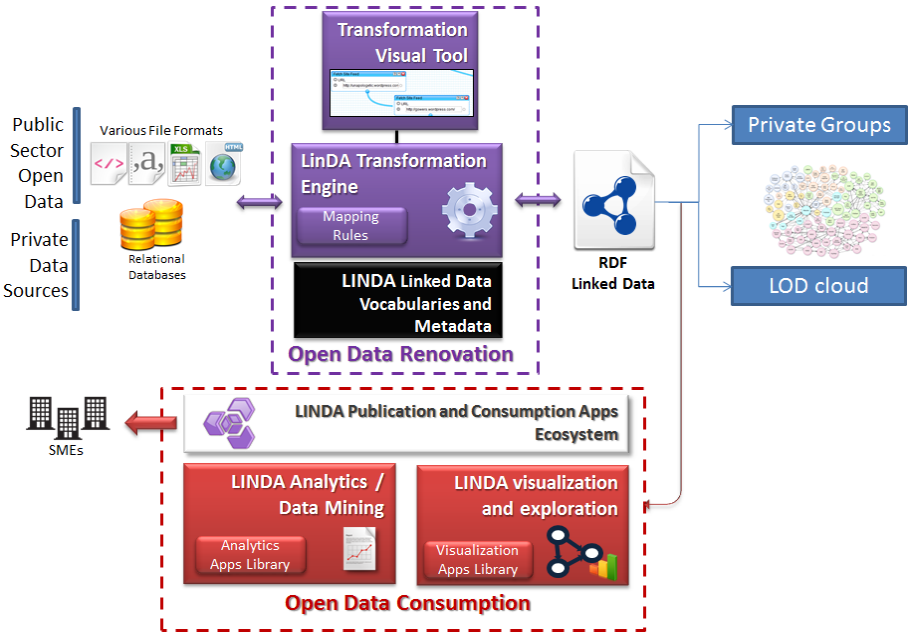

Fig. 1. The LinDA Overall Concept.

2.2 The LinDA Workflow

From a user perspective, the main LinDA workflow can be summarized in three simple steps, as illustrated in Figure 2.

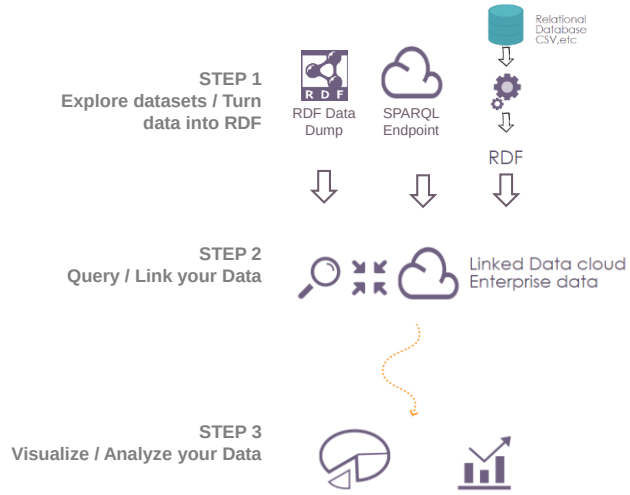

Fig. 2. The LinDA Workflow.

More specifically the three workflow steps are:

- **Step 1 – Explore Datasets/Turn Data into RDF:** Using the LinDA toolset, users can publish their data as linked data in a few, simple steps. In cases where the data are not available in RDF format, the users can simply connect to their database(s), select the data table they want and make their mappings to popular and standardized vocabularies. LinDA assists even more by providing automatic suggestions to the mapping process. Based on the defined mappings, transformation from various formats (e.g. csv, relational database) to RDF is realised.

- **Step 2 - Query/Link your Data:** With the LinDA toolset, users can perform simple or complex queries through an intuitive graphical environment that eliminates the need for SPARQL Protocol and RDF Query Language (SPARQL) syntax. In addition to the submission of queries, interlinking of instances is supported, where the designer lets the end user ignore its instance's data source and handle instances as if they were defined in the same data source. The possible types of interlinking vary according to the interlinking element that is used. More specifically, classes and object/datatype properties can be combined in a versatile way, during the interlinking procedure. Hereinafter, for the sake of homogeneous representation, all interlinking endpoints will be referred as interlinking types. The interlinking of instances of the types [A] and [B] can occur in several ways: (i) instances can be interlinked directly to each other, in which case an entity (URI) is fetched in the query results if belongs to both types [A] and [B] at the same time; (ii) an instance of type [A] can be interlinked to an instance of type [B] via a property, where [A].p is bound to be an instance of type [B] (“owl:sameas" interlinking) and (iii) instances can be interlinked by their properties, where [A].p = [B].q, given that [A].p and [B].q refer to the same URI or that [A].p and [B].q are literals (strings, numbers, dates etc.) with the same value.

- **Step 3 - Visualize/Analyse your Data:** the LinDA toolset can help enterprise users gain insights from the data that the company generates or consumes through the support of a set of visualization and analytics services. LinDA supports visualisations over different categories of data, e.g. statistical, geographical, temporal, arbitrary data, as well as a largely automatic visualization workflow for matching and binding data to visualizations. As far as the analytics services are concerned, they are presented in detail in the following subsection.

According to this workflow, the user can utilize either external public data or internal, private sources. If the initial data source is in RDF format, the user can directly insert the data source to the available data sources of the LinDA Workbench. If the initial data source is in another format (relational database, csv, etc.), the LinDA Workbench guides the user to the toolset responsible for transformations in order to transform the data into the RDF format, with the utilization of popular linked data vocabularies. Once in RDF, the user can then visit the list of data sources and activate one of the available LinDA services. More specifically, the user has the option to a) visualize the selected RDF data source, b) analyse it, c) query it and d) edit/update/delete it.

3 Information Driven Cyber Security Management

3.1 Scenario Description and Implementation at LinDA Workbench

In the examined case, information based on a set of attacks in a small enterprise environment is collected based on the installation of a honeypot. The information regards different type of attacks such as authentication abuse, sql injections etc. The attacks have been recorded for 11 months using raw packet interception mechanisms. Each connection attempt that was classified as malicious was analyzed in a near-time manner. The endpoint-ip of the attacker was submitted to several third-party services in order to infer the location (using a GeoIP resolution service), the size of the originating subclass (using WHOIS services), the possible existence of DNS entries that are associated with the IP (using reverse IP services) and the blacklisting level (using ~40 open lists). In addition each IP was port-scanned and checked for vulnerabilities. The results of this analysis had to be represented in a common format. In order to support common representation of the collected data and support their re-usability and interconnection, a specific cybersecurity oriented ontology is designed. This ontology (see Figure 3) describes the main artefacts of a cyber-attack and specifically:

- the networking environment of the attacker including its IP address, the network size, range and name;
- the hosting environment of the attacker including information regarding the hosting operating system and its vulnerabilities, the open ports detected, the blacklisting level of the considered host based on its IP address, the number of virtual hosts;
- the type of the enterprise/organization where the attack is produced as well as locality information (location of the host including geospatial coordinates);

- the type of the attack based on its classification according to existing cyber-attacks vocabularies, such as CAPEC (https://capec.mitre.org/);
- the date, day and time of the attack taking into account the time zone of the attacking host.

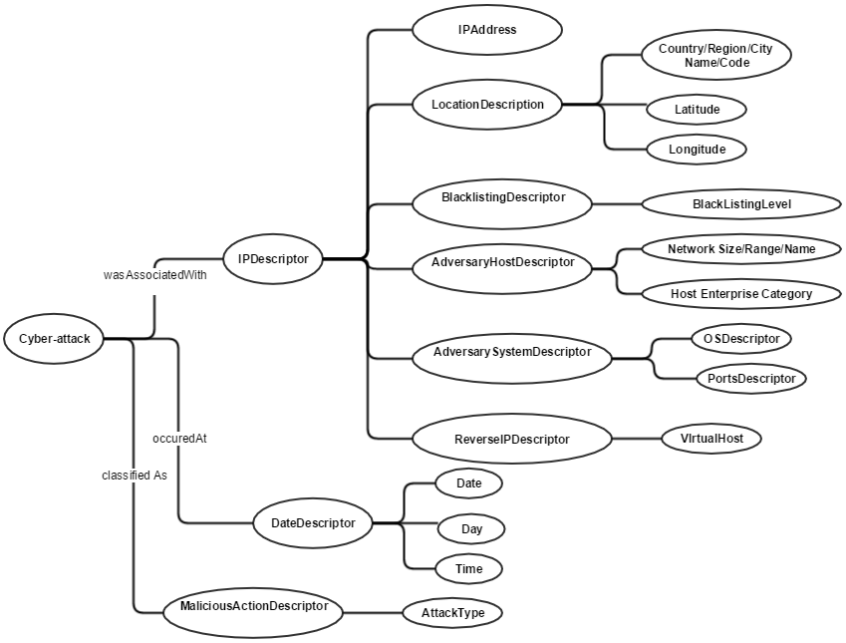

Fig. 3. The Cybersecurity Ontology.

Upon the collection of the cyber-attacks data, their integration in the LinDA Workbench in raw format is taking place. Following, the LinDA Transformation Engine is being used for mapping of the collected raw data into the defined cybersecurity ontology and the production of the RDF data for further analysis. Following, the LinDA Query Designer is being used for design of a set of queries over the available data as well as the definition of possible interlinking of data. An indicative query produced through the Query Designer is depicted in Figure 4.

Next, the produced queries may trigger the initiation of visualization or analytic process through the LinDA visualization tool and the LinDA analytics and data mining tool accordingly. The overall analysis realized is described in the following subsection.

3.2 Analysis Overview

The first step of our analysis regards the production of a set of descriptive statistics, aiming at getting some insights about the available data through monitoring of the variation of selected parameters as well as the production of a set of visualisations.

In Table 1, the number of cyber-attacks from the top 10 countries (in terms of number of cyber-attacks) is detailed, while Figure 5 provides a geomap of the distribution

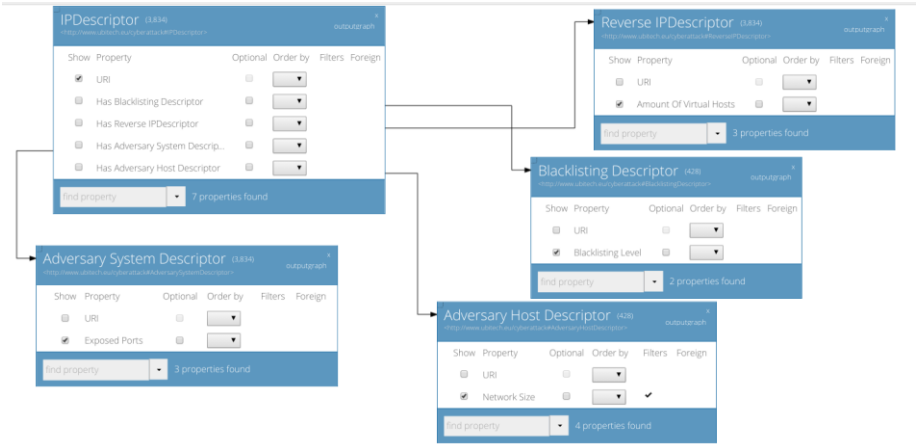

Fig. 4. Indicative Query by LinDA Query Designer.

of the cyber-attacks adversary hosts location. Following, Figure 6 provides a bar plot of the number of attacks per country combined with the average blacklisting index per country, aiming at examining the potential severity of attacks, especially from countries where this number is large (the higher blacklisting index, the more epidemic can be considered an attack).

Table 1. Cyber-attacks per country (top 10).

Country	Number of attacks
China	9715
Hong Kong	2072
Unknown	1577
Malaysia	383
United States	380
Netherlands	138
Germany	61
Republic of Korea	43
Spain	39
India	33

In Table 2, a summary of the number of attacks based on the network range of the network that the IP address of the adversary host is coming from is provided. Larger networks possibly regard home devices that have acquired IP addresses through large telecom operators' networks, while smaller networks may refer to public or private organizations and enterprises that have their own IP address pool.

Following, the cyber-security analyst is interested to have an overview of the trend followed with regards to the daily number of monitored cyber-attacks, aiming at the identification of periodical patterns that could lead to immediate protection actions in the future. Figure 7 is produced for this purpose, depicting the evolution of cyber-attacks monitored in the enterprise's environment for a nine months period.

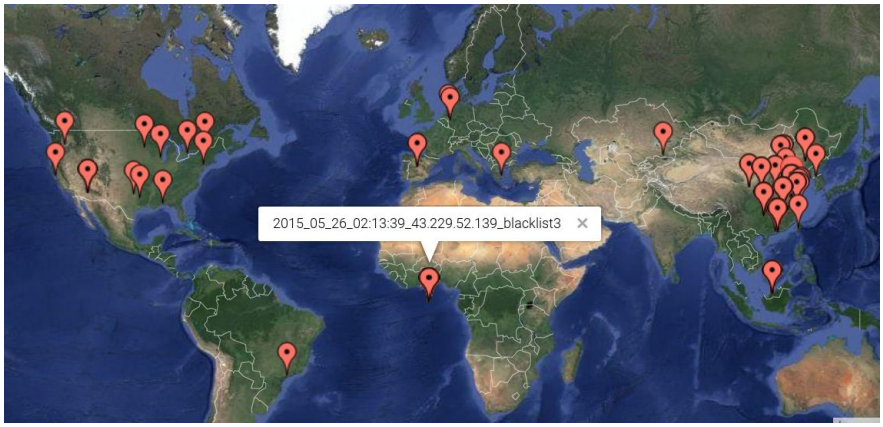

Fig. 5. Geomap of cyber-attacks adversary hosts location.

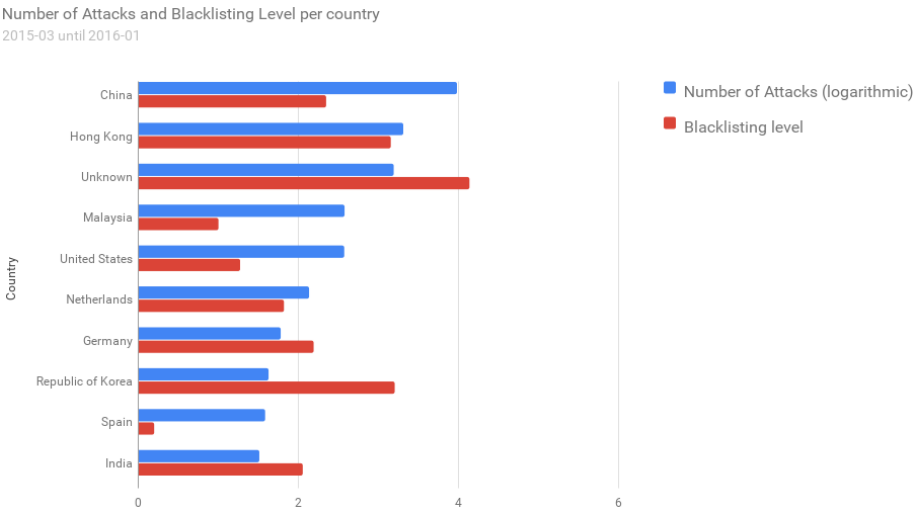

Fig. 6. Number of attacks and blacklisting index per country.

Table 2. Number of attacks per network size.

Netmask	Number of attacks
255.0.0.0	14459
255.255.0.0	308
255.255.255.0	45
255.255.255.240	34

Finally, a clustering analysis is realized over the available data, targeting at the identification of clusters taking into account the variation of the number of virtual hosts, number of exposed ports and blacklisting index parameters. The clustering analysis results are provided in Table 3, while the produced clusters are also depicted in Figure 8. Upon the interpretation of the clustering analysis, it could be argued that cluster 1 regards possibly compromised hosts that are used for botnet-expansion or

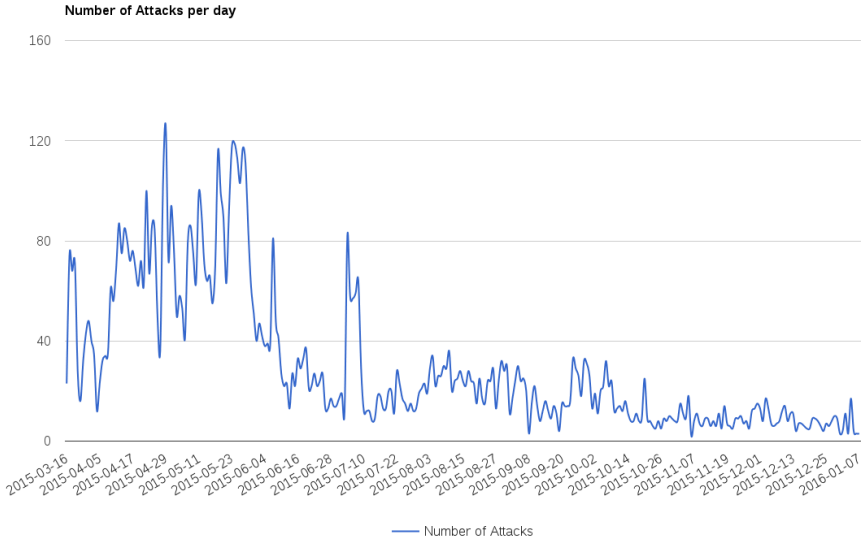

Fig. 7. Variation of number of cyber-attacks during the monitored time period.

Table 3. Clustering Analysis Results.

Cluster	#of virtual hosts	#of exposed ports	#blacklisting index
1	-0.01686813	-0.5667083	-0.6397090
2	-0.02852310	-0.5409678	0.8961126
3	0.07096821	1.7396984	-0.3635785

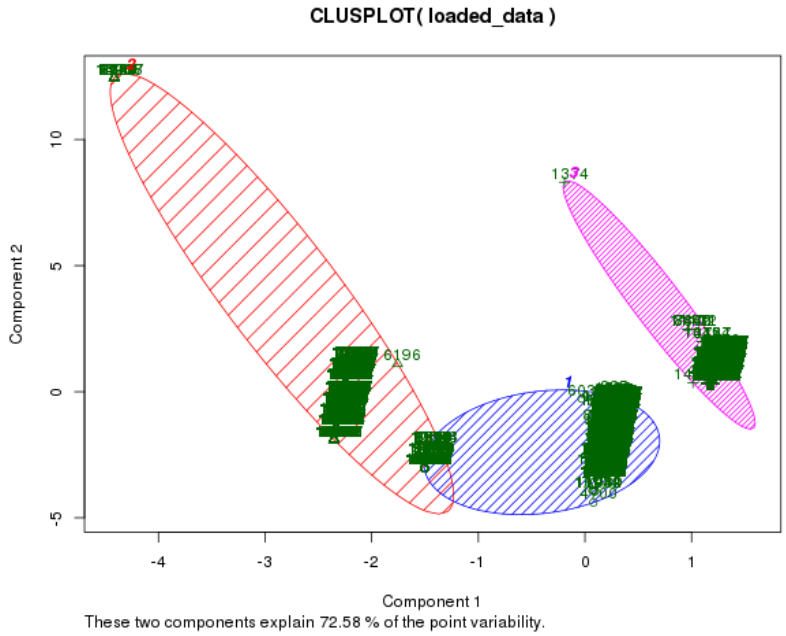

Fig. 8. Variation of number of cyber-attacks during the monitored time period.

dedicated attacks or DDOS attacks, while cluster 2 regards less focused attacks, possibly from mature botnets which the control and command server is not taken down. However, these interpretations have to be further examined and validated based on updated versions of the acquired datasets.

4 Conclusions and Future Work

By taking into account the continuous evolution of cyber threats and the need for design of novel solutions for supporting information driven cybersecurity management, an approach for producing and exploiting linked data from cyber-attacks towards the production of added-value analytics and insights has been provided.

The proposed information driven cyber security management approach aims at enabling effective decision making, threat and risks management through the efficient processing of heterogeneous information flows. The approach is targeting at the provision of a set of information management, analysis and visualisation tools to end users responsible for the deployment, monitoring and management of cyber security solutions (including improved information processing, analysis and, where necessary, exchange functionalities), based on the workbench that is already developed within the framework of the LinDA project.

Prior to the provision of information management functionalities, the approach facilitates the effective collection and harmonization of internal and external information sources related to cyber security management, based on the design of a cyber-attacks representation model. Linked and open data technologies are being used for mapping of the collected information to the developed model and publication/consumption of the available data that can be openly published to commonly used repositories or data that have to be kept in private repositories within the boundaries of an enterprise/public organization.

Based on the deployment and operation of a small-scale scenario upon data collected on a small enterprise environment, indicative analysis is realized leading to a set of insights and validating the efficiency and applicability of the proposed approach. It could be argued that the proposed approach can help enterprises enhancing their experience of managing and processing cyber-attacks data, in ways not available before. It can provide them the potential to produce advanced knowledge, leveraging the power of linked data analytics, for effective information driven cyber security management. However, in order to be able to easily adopt and integrate the usage of such an ecosystem in their daily processes, they have also to take into account the need for an initial learning curve as well as the involvement of data scientists in the specification of the analysis and the interpretation of the analysis results.

With regards to plans for future work, a set of open issues are identified. These include the need for extending the designed cyber security model in order to be more descriptive and applicable to a wider number of cyber threats, the need for interlinking information collected within an enterprise with information openly available in the web for realization of analysis that can lead to more advanced insights and the need for tackling of challenges related to the management of big data and the adoption of a distributed nature of the execution mode.

Acknowledgement. This work has been co-funded by the LinDA project, a European Commission research program under Contract Number FP7-610565.

References

1. Sqrrl report: Linked Data For Cyber Defense – Available Online: http://sqrrl.com/media/linked-data-cyber.pdf
2. Fotopoulou, E., Hasapis, P., Zafeiropoulos, A., Papaspyros, D., Mouzakitis, S. & Zanetti, N., Exploiting Linked Data Towards the Production of Added-Value Business Analytics and Vice-versa, DATA 2015 Conference, Colmar, Alsace, France, 20-22 July 2015.
3. The LinDA Project, Available Online: http://linda-project.eu/
4. The LinDA Workbench, Available Online: http://linda.epu.ntua.gr/
5. The R Project, Available Online: https://www.r-project.org/

An Energy Flexibility Framework on the Internet of Things

Thibaut Le Guilly[1], Laurynas Šikšnys[1], Michele Albano[2], Per D. Pedersen[3], Petr Stluka[4], Luis L. Ferreira[2], Arne Skou[1], Torben Bach Pedersen[1] and Petur Olsen[1]

[1]Aalborg University, Department of Computer Science, Aalborg, Denmark
[2]CISTER/INESC-TEC, ISEP, Porto, Portugal
[3]Neogrid Technologies, Aalborg, Denmark
[4]Honeywell ACS Global Labs, Prague, Czech Republic
{thibaut, siksnys, ask, tbp, petur}@cs.aau.dk, {mialb, llf}@isep.ipp.pt, pdp@neogrid.dk, petr.stluka@honeywell.com

Abstract. This paper presents a framework for management of flexible energy loads in the context of the Internet of Things and the Smart Grid. The framework takes place in the European project Arrowhead, and aims at taking advantage of the flexibility (in time and power) of energy production and consumption offered by sets of devices, appliances or buildings, to help at solving the issue of fluctuating energy production of renewable energies. The underlying concepts are explained, the actors involved in the framework, their incentives and interactions are detailed, and a technical overview is provided. An implementation of the framework is presented, as well as the expected results of the pilots.

1 Introduction

The Internet of Things (IoT) enlarges the Internet to physical objects, extending its usage to various applications such as Smart Grids. Most of these objects are pervasive and mainly interact with other Internet devices such as database servers, other objects or services. With many connected objects[1], using a variety of heterogeneous technologies and protocols, managing their interconnection is a challenge.

Service Oriented Architectures (SOA) have been developed to abstract the specificity of devices and networks and obtain consistent access to functionalities provided by the objects (e.g. [1]). In addition, a lot of effort has been put into filling the syntactic and semantic gaps that exist between networks and applications [2, 3], as demonstrated by the results of the European Connect project [4]. However, a remaining open issue is how to create smooth interconnection between service providers and consumers in the IoT. The Arrowhead project [2] aims at providing a solution to this issue, by developing a framework [5] for IoT applications, including a set of essential services, namely:

- service discovery;

[1] Cisco estimates the number of connected objects to reach 50 billion by 2020.
[2] www.arrowhead.eu

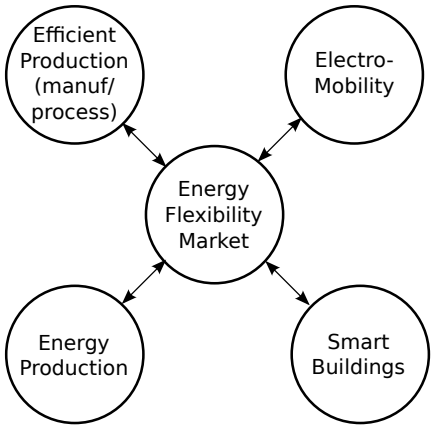

Fig. 1. The five pilot domains of the Arrowhead project.

- authentication;
- authorization.

In fact, if service discovery is sufficient to establish connection between a service consumer and provider, authorization is usually required, in particular for Cyber-Physical Systems (CPS) involving critical components. This is the case in all the pilot domains of the project, which cover the areas or production, electro-mobility, energy production, smart buildings and an energy flexibility market (or flexibility market). These five pilots provide a good sample of the diversity of applications that will be provided by the IoT. As illustrated by Figure 1, the flexibility market is the common denominator between the different pilots, and is expected to provide them its services.

This central position of the energy flexibility market illustrates well the fact that energy will be an important application domain for the IoT, as the challenges that are faced in this area are essential for the future of our society. A European Union directive from 2009 requires the EU countries to fulfill at least 20% of their overall energy needs from renewable sources by 2020 [6]. More recently, Denmark has set a 50% target by 2020 [7]. To attain these objectives, many European projects [8–10] aiming at improving and understanding the production and consumption of electricity have been conducted. Their ambition is to move the current electricity network, the grid, to a *Smart Grid* equipped with smart meters and appliances providing measurement and control capabilities [11].

A main issue with the increasing part of renewable energy sources such as solar cells and wind turbines, is that production from renewable energy sources fluctuates and is not available at all time. It is thus necessary to adjust the behavior of consumers to adapt to the fluctuating production. Solutions to these issues are known as *demand response mechanisms* [12], that provide incentives to end users to modify their consumption behavior to better match the production. A common demand response mechanism is to increase the price in periods of high demand and low production, and reduce it in periods of low demand or high production. Several European programs have been devised and dynamic prices already exist in some countries [13, 14]. Another mechanism, developed in the European project MIRABEL [15], uses the flexibility offered by

some devices and appliances. For example, a Heating, Ventilation and Air Conditioning (HVAC) system can be set with a comfort temperature interval in which it can operate to adjust its consumption pattern. This flexibility can be used to adapt consumption and better match production, or to offload the grid in peak times. However, in order to make use of this flexibility, there is a need for measuring, predicting, and planning consumed and produced energy. We propose in this paper a framework for managing energy flexibility based on so-called FlexOffers, from the end user to a flexibility market where it is traded and assigned an optimal value. The objective is to enable actors of the energy domain to buy flexibility and have more freedom in distributing loads in the grid. The main contribution of this paper is to define the details of this framework, the different actors, their possible interest and their relationships, and to present the underlying ICT infrastructure enabling its deployment. Pilot demonstrations currently taking place in the Arrowhead project are also presented to discuss the applicability of the framework.

The paper is organized as follows. The concept of flexibility and FlexOffer are presented in Section 2. The framework, the actors that compose it and their interactions are detailed in Section 3. An overview of the framework implementation architecture is provided in Section 4. Pilots are presented in Section 5. Finally, related work is discussed in Section 6, and we conclude and discuss future work in Section 7.

2 FlexOffer Concept

This section introduces the concept of FlexOffer, that encodes necessary parameters of flexible loads to facilitate their management. Generation and aggregation of such FlexOffers are then discussed.

2.1 FlexOffer

The flexibility framework presented in this paper is based on the concept of FlexOffer. As already mentioned, this concept was developed in the European MIRABEL project. It provides a way to formally represent flexible energy loads in terms of time and energy, and contains the information necessary to manage them. A visual representation of a FlexOffer in one of its simplest forms is shown in Figure 2. The bars in this graph represent a flexible consumption load from a component with flexible consumption that we name *Flexible Resource* (FR). Each bar represents the consumption for a given time slice. The lower area represents the minimum amount of energy the FR needs to provide its service. The upper area represents an energy interval in which it can change its consumption while ensuring predefined constraints (e.g. temperature comfort). The amount of energy consumed at each slice can thus vary within the interval given by the upper area. Flexibility in terms of energy amounts is referred to as *energy flexibility*. Note that this simple FlexOffer contains only positive flexibility, but some FRs can also contain negative one, representing flexibility in production. The second type of flexibility is time flexibility, and typically occurs when a given load can be shifted in time within a given interval, as illustrated in Figure 2. The time shift is constrained by an *earliest start*, before which the load should not be assigned, and a *latest end* at which it should have been consumed. A baseline is also assigned to each FlexOffer, that

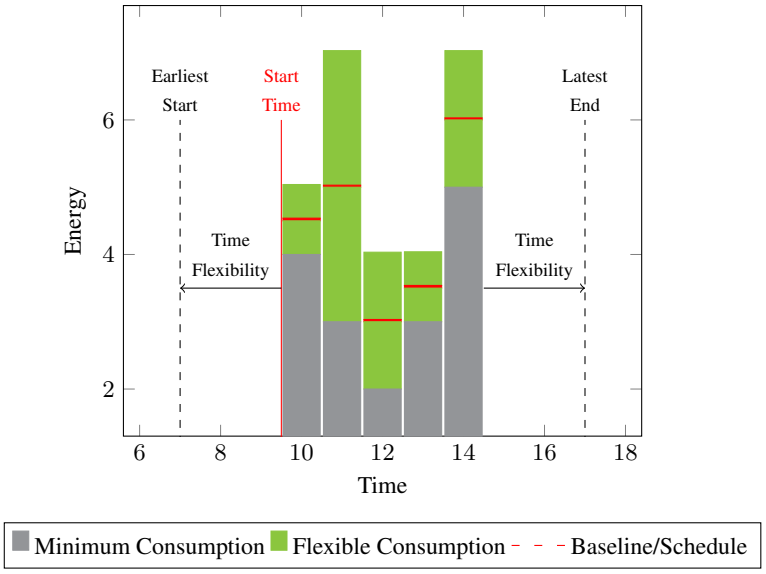

Fig. 2. Example of a FlexOffer.

represents the default consumption plan that the associated FR will follow if it does not receive any update. This baseline schedule can be updated to modify the consumption pattern within the flexibility domain.

2.2 FlexOffer Generation

Generating sound FlexOffers for FRs is not trivial. For FRs that continuously consume energy, such as heat pumps, FlexOffers are typically generated on an hourly or daily basis. Other FRs can emit FlexOffers when needed. Generating a FlexOffer with energy flexibility for FRs implies predicting the consumption (or production) of an FR required to satisfy a given set of constraints, as for example a comfort temperature interval. This is most often done using a model of the FR, its environment including relevant parameters for the predictions such as temperature or solar radiation, and environmental constraints such as comfort settings. Using the model, energy and time flexibility are estimated using various techniques, such as linear programming. Details about generating FlexOffer at the device level can be found in [16].

2.3 FlexOffer Aggregation

FlexOffers most often do not represent large flexible loads. Thus, a single FlexOffer is of little interest to balance energy loads on the grid, where required flexibility is much higher. At the same time, managing large numbers of FlexOffers is tedious and complex. A common solution to facilitate the management of energy loads is to aggregate them. Similarly, FlexOffers can be aggregated into aggregated FlexOffers with larger flexible energy loads. Once an aggregated FlexOffer is assigned a schedule, it needs to

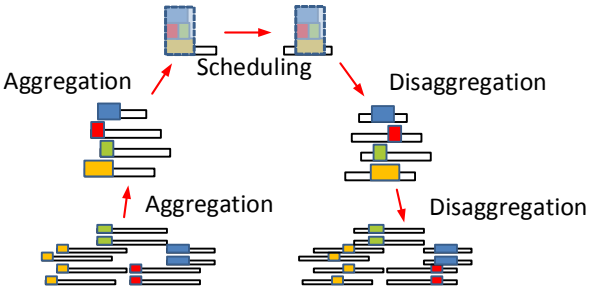

Fig. 3. The workflow of FlexOffer aggregation and FlexOffer schedule disaggregation.

be disaggregated to assign a schedule to each FlexOffer it is composed of. Note that aggregation can be performed multiple times, meaning that an aggregated FlexOffer can contain smaller aggregated FlexOffers, as shown in Figure 3. In general, the flexibility of an aggregated FlexOffer tends to be lower than the sum of the flexibility of the FlexOffers that compose it. This reduction in energy flexibility is however unavoidable to reduce the complexity of flexibility management and the scheduling problem. Note also that aggregating flexible loads is a complex task. To optimize aggregation, FlexOffers can be grouped based on similarity of consumption pattern. More details about aggregation and disaggregation of FlexOffers are provided in [17].

3 Flexibility Framework

An overview of the proposed framework is shown in Figure 4. This section describes its details with the different actors, their interactions and provide examples.

3.1 Flexibility Market

Currently, grid actors trade electricity on existing traditional *day-ahead* (spot), *intra-day*, and *regulation energy markets*. In this Arrowhead pilot, we also consider a so-called flexibility market. It is based on a variation of the product-mix auction [18], in which the commodities are flexible energy loads for specific geographical areas. This model is designed to deal with the "product mix" problem, in which multiple varieties of a product with different costs are supplied, but with a constraint on the total capacity. Here the product is flexibility, and the varieties are positive and negative flexibility. Each bidder can make one or more bids, and each bid contains a set of mutually exclusive offers. Bids in the flexibility market are in fact mutually exclusive, since using both negative and positive flexibility for a given geographical area would not make sense. Two types of bids can be made, *supply bids*, offering flexibility, and *demand bids* requesting it. The auctioneer then selects the market clearing price that for each bid gives bidders the greatest surplus. Offers with negative surplus are rejected. This is visualized in Figure 5. In both graphs, a bid is represented by an horizontal segment. The length of a segment determines the supplied or demanded flexibility amounts, while its position on the Y axis shows the associated price. On the left side, "Up Bids" correspond to bids

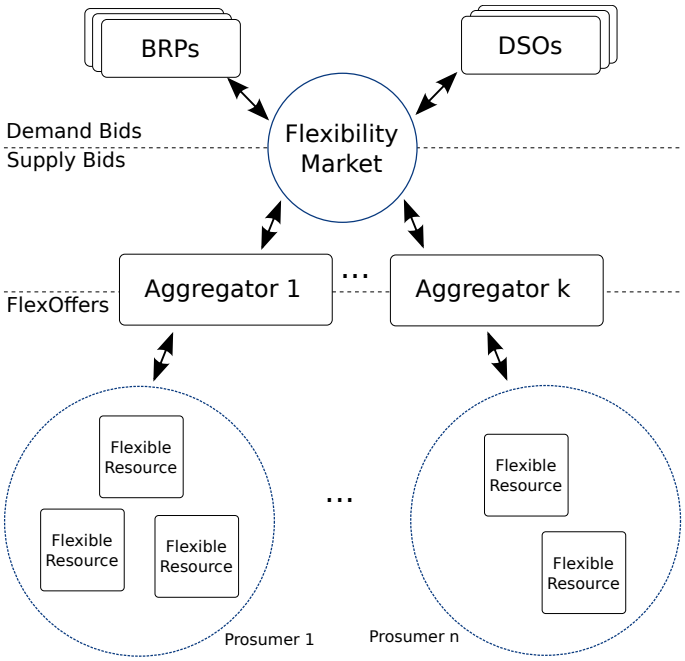

Fig. 4. Overview of the flexibility framework.

for negative flexibility, while "Down Bids" are for positive one. The market clearing price is found at the intersection between demand and supply lines. Supply bids above this line are rejected, similar to demand bids under it. All accepted bids are traded at the market clearing price. In the current implementation of the market, clearing is performed every 15 minutes, but could be adapted to match different needs.

3.2 End Consumer/Producer

The basic elements of the framework are FRs that produce and consume electricity, giving the name *prosumers* to FR owners. A first example is a household, in which we can identify a number of FRs. The Heating, Ventilation and Air Conditioning (HVAC) system is a first important one. If a user agrees to let such systems be controlled flexibly, meaning setting intervals for the different comfort settings, it is possible to generate useful FlexOffers from them. Generating FlexOffers from this type of system is the objective of one of the pilots of the Arrowhead project, which will be discussed in Section 5.1. Similarly, fridges and freezers can be operated in a given interval to offer flexibility. Another type of system that can offer flexibility is an appliance such as a washing machine, tumble dryer or dishwasher. In general, such appliances follow a fixed consumption pattern, that could thus only be shifted in time. This could be achieved by asking users to specify an interval during which they should operate, or a deadline by which a given operation should be terminated. However, at the moment, few of these appliances provide remote control capabilities, making it difficult to explore the applicability and acceptance from users.

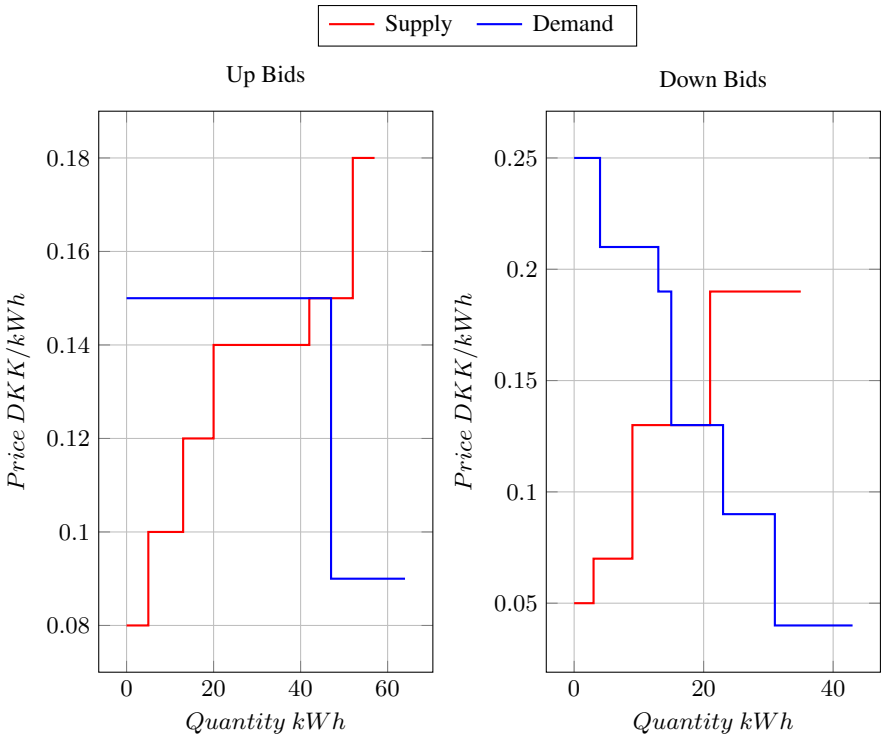

Fig. 5. Visualization of market clearing price.

Electric cars are also interesting, since their numbers is expected to increase in the near future. In fact, they can be charged in a very flexible manner, and can even be used as storage facilities. In addition, they are a real issue for existing grid infrastructures that in some cases will not support convoluted charges. It is also important that user constraints can be enforced, such as ensuring a minimal charge for emergency situations. A pilot on this topic is expected to take place in the context of the Arrowhead project.

Nowadays, houses are getting more equipped with production units such as Photo-Voltaic panels or wind turbines. These can be used to increase the flexibility offered by equipped houses. These units thus provide negative flexibility. Combining information from producing and consuming units makes it possible to generate FlexOffers with large amounts of both positive (consumption) and negative (production) flexibility. Finally, as local storage units are becoming more affordable, they could also be of interest to the framework.

Another type of FRs are public buildings that are essentially equipped with similar type of devices than houses, but with larger capacities, both in terms of production and consumption. The project includes two pilots with buildings equipped with innovative technologies that can be used to generate FlexOffers, and will be discussed in Section 5.

The last important type of prosumer are industrial actors. Industrial processes are in fact using large amounts of energy for manufacturing goods. The consumption patterns of these processes can in some cases be adjusted, by shifting production schedules

or throughput. Industrial actors need to play an important role in the green transition, and tools such as FlexOffers can facilitate their integration to the Smart Grid. In the framework, only this type of prosumers are expected to participate in the flexibility market. Smaller ones will interact with it through an Aggregator.

3.3 Aggregator

An aggregator is a business entity that makes money by aggregating FlexOffers from several FRs and selling them on the flexibility market. It essentially acts as a Commercial Virtual Power Plant (CVPP), providing load-shifting options and (near) real-time control of many FRs on the energy market. As already mentioned, small Prosumers, as for example a household, do not provide large enough flexible loads to be of interest for a market. An Aggregator thus makes a contract with a number of these Prosumers, giving it the right to control their FRs based on the FlexOffers they emit. In exchange, it must reward the offered and used flexibility with a previously agreed price scheme. We propose that the actual reward be calculated based on:

- The number of FlexOffers issued by a Prosumer;
- The amount of flexibility offered by the prosumer, both in time and energy amounts;
- The amount of flexibility used by the Aggregator to balance loads in the grid;
- Other parameters such as number of actual activations, accuracy with which schedules are followed, etc.

Once a contract is agreed upon between an Aggregator and a Prosumer, the generation, aggregation and schedule of energy loads can start. The interaction between an Aggregator and an FR is shown in Figure 3.3. When an FR sends a FlexOffer to an Aggregator, the first step the Aggregator takes is to check for validity. Essentially this means ensuring that the time intervals of the FlexOffer are consistent. It then decides, based on the state of the grid and other parameters, if the FlexOffer is useful for its needs. If it is, the Aggregator accepts it, notifies the FR and aggregates the received FlexOffer with other ones, to produce one or more *aggregated* FlexOffers. During planning, the Aggregator can update the baseline of the FlexOffer by assigning it a schedule. Scheduling can be performed multiple times to react to planning changes, up to a time included in the FlexOffer. After this time, the FlexOffer is executed by the FR, meaning that it consumes (or produces) electricity following the assigned schedule. The consumption (or production) is measured to ensure that the schedule is respected by the FR. In the billing phase (every month), the Aggregator computes all of its revenues, losses and bills.

In the planning phase, an Aggregator uses the pool of aggregated FlexOffers to generate bids for the flexibility market. For each aggregated FlexOffer, it sends one supply bid with two offers. One for positive flexibility and one for negative. Recall that among these two offers only one can be accepted. Winning offers result in assignments of schedules to corresponding aggregated FlexOffer. The Aggregator can also trade on other existing power markets (e.g., ELSPOT or ELBAS) and enter into bilateral agreements with other parties such as Balance Responsible Parties. The Aggregator business model is shown in Figure 6.

Aggregators can also be specialized based on the type of FRs they handle. As already mentioned, aggregation can be optimized by grouping similar FlexOffers.

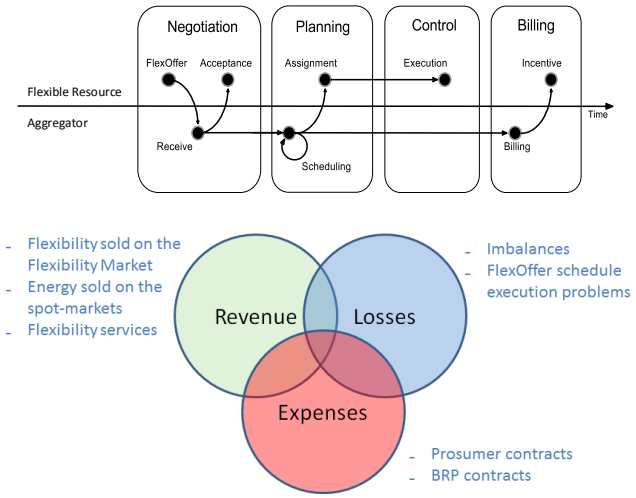

Fig. 6. Aggregator business model.

3.4 Distribution System Operators

A Distribution System Operator (DSO), is responsible for operating, maintaining, and when necessary developing the distribution system in a given area, delivering (and possibly receiving) electricity to (from) end users. This is in contrast to a Transmission System Operator (TSO) that have similar responsibilities for the transmission system, delivering electricity from large generation units to industrial consumers and local transformers. In Denmark for example, DSOs operate under 65kV while TSOs operate above.

The increasing number of energy consuming devices and appliances, including heat pumps and electric cars, lead to an increase of the load on the distribution grid. The addition of Distributed Energy Resources (DER) connected to it such as wind turbines puts even more pressure on existing installations. However, most issues arise during peak periods. To solve them, DSOs thus have two options:

- Strengthen the grid or
- Smoothen the load.

Strengthening the grid requires large investments from DSOs. Smoothening the load is more cost effective and can be used in addition to strengthening the grid. The flexibility market can be used by DSOs to that effect. They access the market by expressing interest in flexibility for specific geographical areas in which they operate. For example a DSO can emit a bid expressing interest in positive flexibility in a given area, to reduce congestion points. The interaction between DSOs and the energy market is as follows.

Step 1. Forecast of the loads on the grid and identification of possible bottlenecks. The baseline loads of Aggregators are queried and used to improve the accuracy of forecasting and congestion detection.

Step 2. During market opening time, for each bottle-neck point, a bid with several offers is generated, with different demands for flexibility at different prices. Each offer can for example represent a potential solution for a bottle-neck in the grid.
Step 3. If an offer is accepted, the DSO enters into an agreement with the wining party (for example an Aggregator) to make use of its flexibility.
Step 4. The DSO updates the load forecasts to mirror the deviations of the winning bids and re-computes bottle-necks.

Alternatively, or additionally, they can also make bilateral agreements with large prosumers to exclusively handle their FlexOffers.

3.5 Balance Responsible Party

Balance Responsible Party (BRP) is a delegated role from a TSO to ensure balance in the transmission grid, e.g. ensuring fitness between consumption and production. Today this is done by trading on existing energy markets, based on expected production and consumption. The main task of the BRP is to predict hourly energy flow up to 36 hours ahead and trade electricity accordingly. However, with fluctuating energy sources like wind turbines and PVs this is becoming more difficult. The interest from the BRP in a flexibility market is to "buy" flexibility from industrial FRs and Aggregators to balance power within their respective grid area. Interaction between BRPs and the market place is similar to the DSOs interaction.

4 Software Architecture Overview

The flexibility framework introduced in the previous section, to be put in practice, is supported by a number of software components and ICT solutions enabling information exchange, FlexOffer generation, aggregation and control of FRs. These components are implemented in Java, following a set of pre-defined interfaces. They are supported by the Arrowhead core services, a set of management services designed to facilitate the deployment of IoT applications. It aims at facilitating interconnection between systems of different application domains (e.g.: industrial automation, airplane maintenance, energy production, home automation, smart grids). The objective is to enable cross-domain applications, among which energy is a central point. Figure 7 illustrates the different system components as well and their interconnections, that will be described in this section.

4.1 Arrowhead Framework

The Arrowhead core services facilitate interactions between the different systems implementing the framework. The Service Registry service allows service providers to advertise their services, and service consumers to look them up.

The Orchestration service allows to automatically connect a service provider to a service consumer. This is based on a query from the latter, containing requirements that the service should satisfy. The Orchestrator essentially performs a look up of the

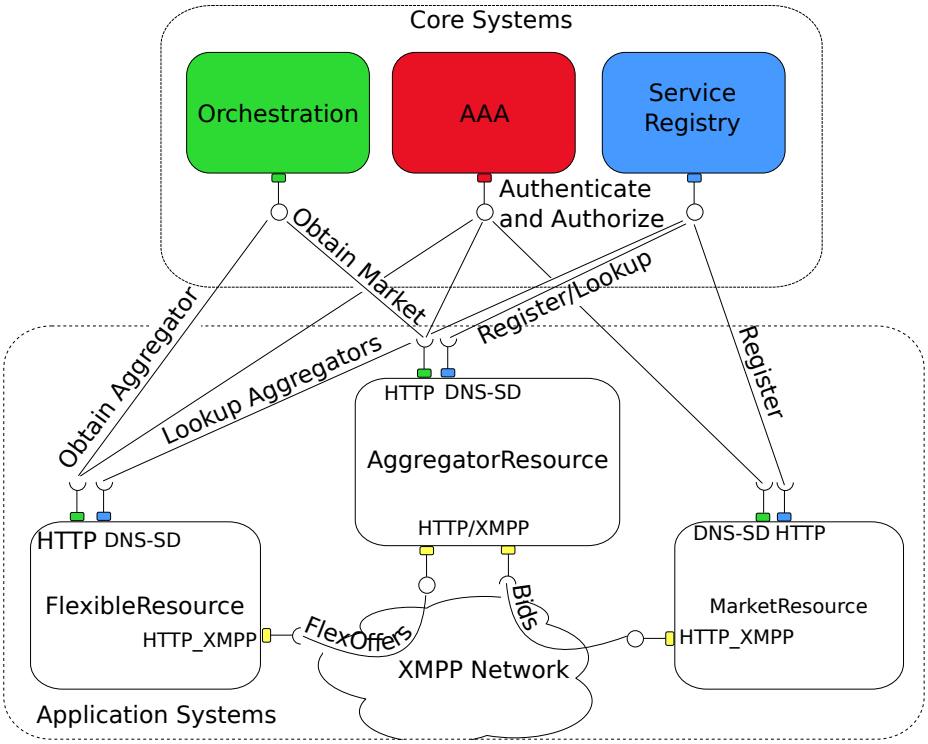

Fig. 7. Software architecture of the Virtual Market of Energy.

information stored in the Service Registry, to find the best service to satisfy a given query. Note that orchestration can return a simple service (e.g.: an AggregatorResource specialized into heat pumps) or a composed service (e.g.: a set of MarketResources to be used in round robin, or at different time of the day).

The Authentication, Authorization and Accounting (AAA) service enables verification of component identities, access control to functionalities and accountability measure in case of misuse. This is essential in the framework as malicious actions could lead to damages to grid infrastructure. As an example, criminals could impact grid operations by misbehaving in the framework, or taking financial benefit by claiming dishonest information.

The Arrowhead core services were instrumental in easing the implementation of the flexibility framework, and to ensure a good level of performance and security for the interacting systems, by providing a Service Oriented platform to support all the interactions for the functioning of the systems at hand.

4.2 Market Resource

The MarketResource component encapsulates functionalities of the flexibility market. The architecture allows for multiple markets to be defined, which could correspond to different geographical areas or type of flexibility traded. After authentication, the

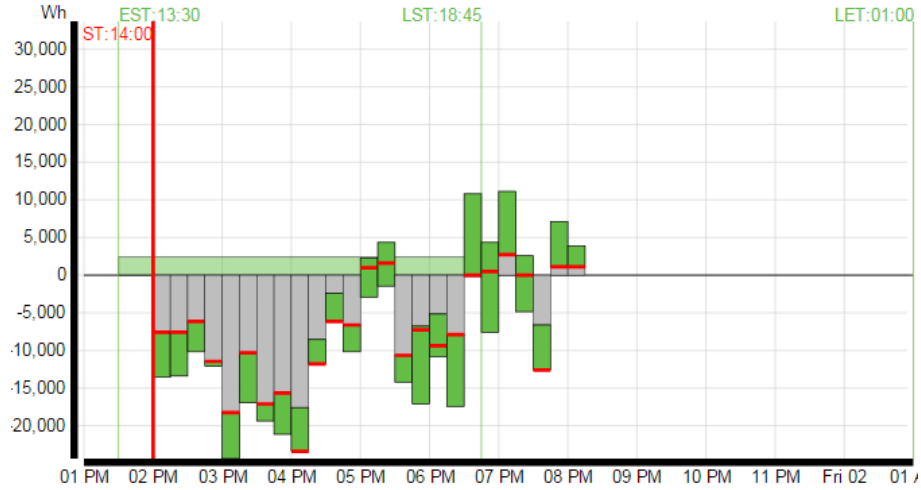

Fig. 8. FlexOffer visualization in an FR graphical user interface.

market makes its interface available to authorized bidders. It then conducts the auction as already described in Section 3.1.

The current implementation of the Market Resource provides a graphical user interface enabling its management and monitoring. It consists of a web application and includes for example the visualization of market clearings shown in Figure 5.

4.3 Flexible Resource

As already mentioned, FRs are implemented in a software component of the same name. The functionalities handled by the FR component are:

FlexOffer Generation. Sending FlexOffers to an Aggregator;
Consumption Management. Following assigned consumption schedule.

To enter the framework, an FR needs to authenticate to the AAA service, and be an authorized entity. It can then look up an adequate Aggregator with respect to its flexibility and geographical area using Service Registry or Orchestration services. The returned information enables it to connect to an Aggregator, if authorized by this one. This means that there exists an agreed contract between Aggregator and FR, as mentioned in Section 3.3.

The currently implemented FR components generally provide two user interfaces (UIs), mostly through web applications. A first one is used to set configuration parameters for FRs, such as user constraints. Examples will be shown in Section 5. A second UI is used to monitor flexibility of the system, and is common to all FRs. It contains for example a visual representation of FlexOffers, illustrated by Figure 8, or information on generated FlexOffers and rewards as shown in Table 1.

Table 1. Example of information about generated FlexOffer provided in FR graphical user interfaces.

Item	Value	Price
Number of FlexOffers	20	
Fixed reward for providing flexibility		10 DKK
Total Time Flexibility	289 time units (15min.)	28.90 DKK
Total Energy Flexibility	409,137.63 Wh	40.91 DKK
Number of baseline updates	3	15 DKK
Used time flexibility	3 time units (15min.)	9 DKK
Used energy flexibility	10,232.91 Wh	21.44 DKK

4.4 Aggregator Resource

The Aggregator Resource component implements the functionalities offered by an Aggregator. It acts as a service provider towards FRs, enabling them to submit generated FlexOffers, informing them on their status, and sending back consumption schedules when FRs baseline consumption patterns are modified through wining bids and disaggregation. After authenticating, it advertises its service to the Service Registry so that it can be found by relevant FRs. If multiple markets are available, it can also use Orchestration or Service Registry to find the most relevant one for the flexibility it has to offer. It also uses AAA services to ensure that only authorized FRs can connect to it. Finally, Aggregators offer UIs similarly to FRs, offering visualization and management of FlexOffers, contracts and price information.

4.5 Communication Infrastructure

Interaction between the different components is implemented using different approaches based on communication requirements and component specificities.

Arrowhead Core Services are generally provided over the HTTP protocol. An exception is the Service Discovery that uses DNS-based Service Discovery [19].

Web Interfaces are also provided by each component over HTTP. This makes it easy to access them from any Web compatible device.

Framework Components communicate using HTTP over XMPP [20]. The reason for this choice is that as HTTP is already used for communication with the Arrowhead core services and web applications, reusing it for component communication enables reuse of interfaces, and consistent error handling. However, HTTP makes it difficult to establish two way communications. This is often in houses of buildings where FRs are located, due to Local Area Networks (LAN) and firewalls that prevent incoming connections to networked devices. Using XMPP as an underlying communication layer makes this possible, in addition to enforcing communication encryption. In addition, XMPP is being considered as a transport method for the second version of the Open Automated Demand Response (OpenADR), and the ISO/IEC/IEEE 21451-1-4 [21] standards.

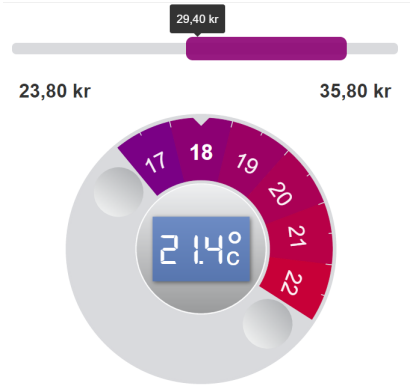

Fig. 9. Screenshot of the application to set interval comfort temperature and obtain visualization of corresponding reward.

5 Pilots

This section introduces three pilots of the Arrowhead project that are currently being developed to explore the applicability of the flexibility framework in different use cases.

5.1 Heat Pumps

The first pilot consists of an individual control of heat pumps installed in occupied residential houses. Each household is provided access to a web application through an FR component, enabling setting of comfort temperatures and visualize associated reward, as shown in Figure 9.

The idea behind the pilot is that a heat pump can be controlled flexibly both in time and energy consumption, while ensuring user constraints such as comfortable temperature interval of the house. The process used in this pilot is as follows:

- Create a model which can predict the energy demand of the house;
- Calculate day-ahead the cheapest energy plan to secure comfort using spot-price;
- Every 15 minute issue a FlexOffer describing the options for decreasing/increasing power consumption;
- Wait for an eventual FlexOffer schedule.

The data used for modelling are historical data for:

- Delivered heat in the house;
- Used energy for hot water;
- Indoor temperature;
- Power consumption of the heat pump;
- Weather data.

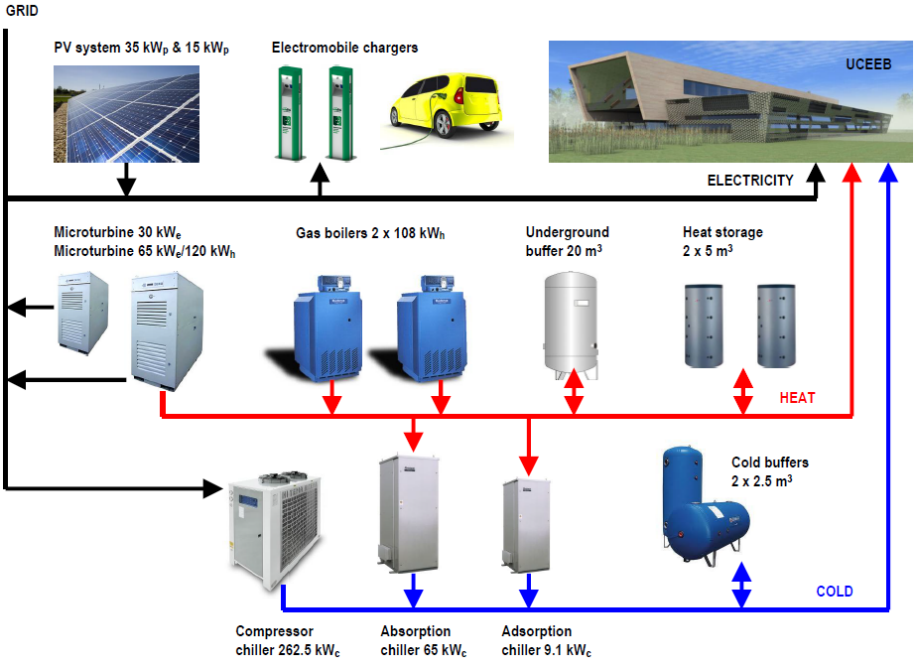

Fig. 10. Overview of the energy grids and resources in the UCEEB building.

Forecasting data are:

- The model;
- Weather forecast;
- User comfort criteria.

To apply assigned consumption schedules, heat pumps are operated via a relay, that can be used to stop them. However, it is not possible to force a heat pump to run.

5.2 Load Management in a University Building

The second pilot is being implemented in the new building of the University Center for Energy Efficient Buildings (UCEEB). It is located in Bůstěhrad, a small town near Prague, Czech Republic. The UCEEB building serves as a complex experimental platform for all research fields related to the area of energy efficient buildings. It integrates a variety of spaces, including open space office, smaller office rooms, large halls and laboratories. It is also an experimental facility with multiple local energy sources integrated into respective distribution grids. As shown in Figure 10, it includes three energy grids, one for electricity, one for heat and one for cold. Energy resources include:

- two photo voltaic fields producing electricity (35 kWp and 12 kWp),
- a combined heat and power (CHP) unit producing heat and electricity,
- two charging stations for electric cars,
- storage units for heat and cold,
- two gas boilers, and
- two chillers.

Prior to the implementation of the FlexOffer concept, a simulation study was conducted to assess suitable scenarios, and determine how the building system could operate in combination with FlexOffers:

- Adjusting the HVAC system set points based on FlexOffers, ensuring the comfort temperature while maximizing the economic benefit for supplying flexibility to the market;
- Enable generation of FlexOffers by manipulating the temperature of hot and cold water;
- Using strategies such as *dynamic pre-cooling* or *pre-heating* to increase the flexibility of the HVAC system during periods of peak consumptions.

Following this study, several rooms with associated electric heaters and fan-coil units were selected for the pilot implementation so that experimentation could be conducted both during hot and cold seasons. An important part of the pilot is a software application connected to the Building Management System (BMS) and a database containing historical data. It is used by the building operator to control the process of FlexOffer generation. This human supervision is essential because specific adjustments of HVAC system operations can potentially lead to uncomfortable situations for the occupants of the building. The building operator is able to balance between comfort level and economic benefits. When the operator interacts with the application, the following functions are provided:

Precool: specifies a time interval in which pre-cooling of the building is allowed;
Duty Cycle: enforces a specified cycling pattern on the functioning of the HVAC system, which can be used to provide flexibility;
Full Flexibility: leaves full control of the system to the flexibility framework for a given time period, thus ignoring any comfort setting;
Adjust Set Points: allows manipulations of room temperatures in the building, but in this case the flexibility is specified by the operator himself.

The objective of this pilot is to experiment with the application of FlexOffers inline with the existing control system of the building, to assess potential benefits and application constraints. Some initial learnings are summarized here:

- The FlexOffer concept presents similarities with applications of Automated Demand Response (ADR) in commercial buildings. One possible difference is that today ADR always triggers a firmly defined load shedding strategy, while FlexOffers are assuming more degrees of freedom in building operation. For this reason it seems important to have a human operator responsible for assessing of how much flexibility may be offered under given conditions, and thus, supervising the overall process of generating flex-offers.
- Operating the HVAC system flexibly implies compromising between maximizing economic benefits and energy savings and overall comfort constraints in the building. For this reason only relatively short alterations of the default control strategy are desirable. This applies primarily to the strategies related to duty cycling and set point adjustments, which both should not span long time intervals. An example of a load shedding strategy implemented using FlexOffers, lasting over 4 hours is

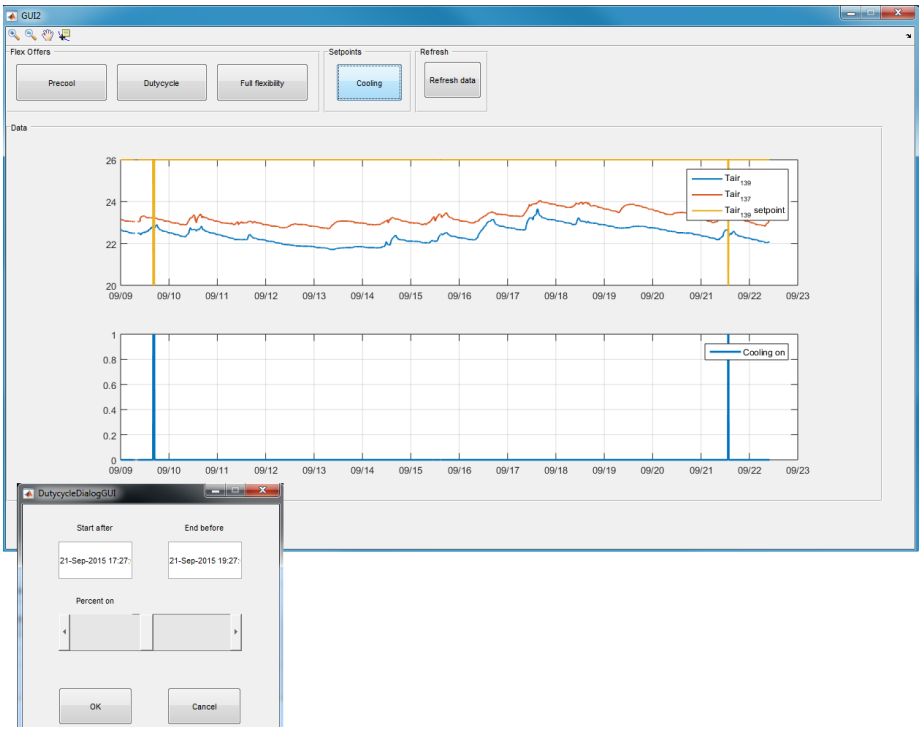

Fig. 11. User Interface used by building operators to specify flexibility parameters.

shown in Figure 12, which shows a systematic reduction compared to the estimated baseline.

– The economic benefits of FlexOffers need to be further studied. A first reason is the so-called rebound effect, already known from ADR projects. It occurs when occupants temporarily increase their heating/cooling demands after the completion of a schedule inducing reduction of comfort level to compensate the possibly experienced discomfort. This can increase energy consumption and operational costs and reduce the overall economic benefits. However, FlexOffers spanning an entire day, with relational constraints between time slices, could be used in the optimization to take into account such effects and help to determine their financial cost.
– Finally, it is important that building owners participating in the flexibility framework receive enough incentives to maintain their interest.

5.3 PVCC

PhotoVoltaic Comfort Cooling (PVCC) is a Danish project that aims at combining technologies with an innovative energy management system to improve the current cooling system of a bank building in Hadsund, Denmark. An overview of the setting is shown in Figure 13. The system is composed of:

Photovoltaic Panels: (PVs) producing electricity from solar energy.

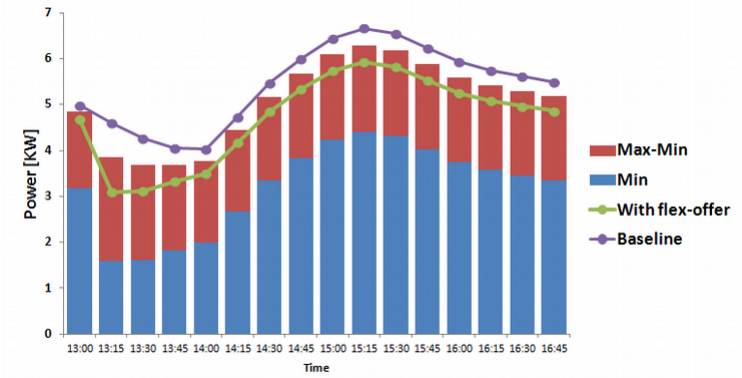

Fig. 12. Example of load shedding implemented using FlexOffers compared to estimated baseline.

A Heatpump: converting the electricity produced by PVs into cool air.
An Ice Bank: that can be used to store thermal energy in the form of ice.
A HVAC: used to control indoor climate.
A Controller: monitoring the different components and control energy production from the PVs and consumption from the heat pump. Its objective is to ensure system stability and improve energy consumption while maintaining a comfortable climate for building workers.

The controller receives information from the Danish Meteorological Institute to better anticipate PV production, thermal changes and optimize energy consumption. It can also receive external control commands from remote clients through the Internet. The system has been deployed for more than a year, and has drastically improved the comfort level of the building. In fact, due to its outside being composed mainly of glass, the sun heating it rapidly increased indoor temperature in summer. Due to the presence of both consuming and producing components, as well as energy storage, this pilot is now being investigated to generated interesting FlexOffers from it, that could be traded on the flexibility market.

6 Related Work

There has been a number of projects exploring the use of flexible loads to solve balancing issues on the grid. Already mentionned, the MIRABEL project [15] introduced the notion of FlexOffer reused in the presented framework. A similar flexibility market has previously been proposed in the iPower project [22]. It provides an overview of the possible interactions between DSOs and Aggregator, detailing their interaction process using a market and contracts to ensure application of the assigned schedule. Here we have provided a more general overview of the components and actors that constitute our proposed framework. Our approach also differs by the use of the FlexOffer concept, as well as the application and implementation of the product-mix auction in the market. In addition, we have provided details about concrete implementations of an ICT infrastructure enabling the deployment of such a market.

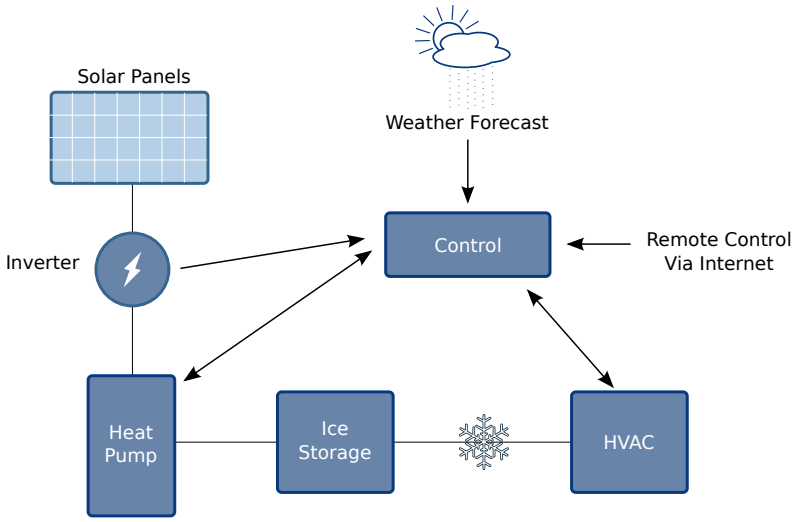

Fig. 13. Overview of the PVCC pilot.

Powermatcher [23] is also an auction based framework for energy balancing using load shifting. FRs are referred to as device agents while Aggregators are referred to as concentrators. It provides an open source library to implement the framework using web sockets over HTTP for communication. Here the integration with the Arrowhead core services and the use of XMPP aims at providing necessary services for facilitating interconnection of the components and security and accountability measures. The CITIES project [24] explores energy flexibility at the city level, and has shown interest in the presented framework.

There is in general active research in the area of energy flexibility. Neupane et al. evaluate the value of flexibility on current regulation markets. Valsomatzis et al. [25] discuss how to compare energy flexibility, a useful technique for managing FlexOffers at the FR or Aggregator level.

Finally, a technical overview of the framework was presented in [8]. Here we have presented a more general overview of the framework, and more details about the flexibility market.

7 Conclusion and Future Work

In this chapter we have presented a framework to leverage flexible energy loads from consuming and producing devices, aggregate them and make them available on a flexibility market for interested parties. We have described the different actors of the framework and detailed their interactions and interest in it. The software components and ICT infrastructure enabling the deployment of the framework were also presented. This includes interaction with the Arrowhead core services, that facilitate integration into the Internet of Things. Finally, we have presented three pilots where the presented framework is currently experimented, that provides a perspective of its possible application in concrete scenarios.

Further work will first consist in consolidating the framework, trying to converge to a standardized approach to trading flexibility and finding synergies with similar approaches. This also includes the ongoing standardization process of the FlexOffer concept, as well as in the interaction between the actors, in collaboration with industrial partners to ensure feasibility of the approaches. The underlying ICT infrastructure is still ongoing further development, with a goal to release an open implementation of the framework providing all necessary services. Research in generation, aggregation and improvement of the flexibility market is also expected to continue to increase flexibility offered by FRs and Aggregators, thus strengthening the interest of the framework. Finally, the flexibility market will also be improved, both from a theoretical and application point of view.

Acknowledgements. This research was supported by the European project Arrowhead.

References

1. Le Guilly, T., Olsen, P., Ravn, A., Rosenkilde, J., Skou, A.: Homeport: Middleware for heterogeneous home automation networks. In: Pervasive Computing and Communications Workshops (PERCOM Workshops), 2013 IEEE International Conference on. (2013) 627–633
2. Issarny, V., Bennaceur, A., Bromberg, Y. D.: Middleware-layer connector synthesis: Beyond state of the art in middleware interoperability. In Bernardo, M., Issarny, V., eds.: Formal Methods for Eternal Networked Software Systems. Volume 6659 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2011) 217–255
3. Blair, G. S., Paolucci, M., Grace, P., Georgantas, N.: Interoperability in complex distributed systems. In Bernardo, M., Issarny, V., eds.: Formal Methods for Eternal Networked Software Systems. Volume 6659 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2011) 1–26
4. Issarny, V., Steffen, B., Jonsson, B., Blair, G., Grace, P., Kwiatkowska, M., Calinescu, R., Inverardi, P., Tivoli, M., Bertolino, A., Sabetta, A.: CONNECT challenges: Towards emergent connectors for eternal networked systems. In: Engineering of Complex Computer Systems, 2009 14th IEEE International Conference on. (2009) 154–161
5. Blomstedt, F., Ferreira, L., Klisics, M., Chrysoulas, C., Martinez de Soria, I., Morin, B., Zabasta, A., Eliasson, J., Johansson, M., Varga, P.: The Arrowhead approach for SOA application development and documentation. In: Industrial Electronics Society, IECON 2014 - 40th Annual Conference of the IEEE. (2014) 2631–2637
6. European Parliament and Council of the European Union: Directive 2009/28/ec of the European parliament and of the council of 23 april 2009 on the promotion of the use of energy from renewable sources and amending and subsequently repealing directives 2001/77/ec and 2003/30/ec. Official Journal of the European Union 52 (2009) 16–62
7. The Danish Ministry of Climate, Energy and Building: Energy policy report (2013)
8. Albano, M., Ferreira, L., Le Guilly, T., Ramiro, M., Faria, J., Perez Duenas, L., Ferreira, R., Gaylard, E., Jorquera Cubas, D., Roarke, E., Lux, D., Scalari, S., Majlund Sorensen, S., Gangolells, M., Pinho, L., Skou, A.: The encourage ict architecture for heterogeneous smart grids. In: EUROCON, 2013 IEEE. (2013) 1383–1390
9. Pedersen, T., Ravn, A., Skou, A.: INTrEPID: A project on energy optimization in buildings. In: Wireless Communications, Vehicular Technology, Information Theory and Aerospace Electronic Systems (VITAE), 2014 4th International Conference on. (2014) 1–4

10. Catalin Felix, C., Mircea, A., Julija, V., Anna, M., Gianluca, F., Eleftherios, A., Manuel Sanchez, J., Constantina, F.: Smart grid projects outlook 2014. Technical report, European Comission (2014)
11. Albano, M., Ferreira, L., Pinho, L.: Convergence of smart grid ICT architectures for the last mile. Industrial Informatics, IEEE Transactions on 11 (2015) 187–197
12. Balijepalli, V., Pradhan, V., Khaparde, S., Shereef, R.: Review of demand response under smart grid paradigm. In: Innovative Smart Grid Technologies - India (ISGT India), 2011 IEEE PES. (2011) 236–243
13. Torriti, J., Hassan, M. G., Leach, M.: Demand response experience in europe: Policies, programmes and implementation. Energy 35 (2010) 1575 – 1583
14. Teixeira, C., et al.: Convergence to the European energy policy in European countries: case studies and comparison. Journal of Social Technologies 4 (2014) 7–24
15. Boehm, M., Dannecker, L., Doms, A., Dovgan, E., Filipič, B., Fischer, U., Lehner, W., Pedersen, T. B., Pitarch, Y., Šikšnys, L., Tušar, T.: Data management in the MIRABEL smart grid system. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops. EDBT-ICDT '12, New York, NY, USA, ACM (2012) 95–102
16. Neupane, B., Pedersen, T., Thiesson, B.: Towards flexibility detection in device-level energy consumption. In Woon, W.L., Aung, Z., Madnick, S., eds.: Data Analytics for Renewable Energy Integration. Volume 8817 of Lecture Notes in Computer Science. Springer International Publishing (2014) 1–16
17. Siksnys, L., Valsomatzis, E., Hose, K., Pedersen, T.: Aggregating and disaggregating flexibility objects. Knowledge and Data Engineering, IEEE Transactions on 27 (2015) 2893–2906
18. Klemperer, P.: The product-mix auction: a new auction design for differentiated goods. Journal of the European Economic Association 8 (2010) 526–536
19. Cheshire, S., Krochmal, M.: DNS-Based Service Discovery. RFC 6763 (2013)
20. Waher, P.: HTTP over XMPP transport. XEP 0332, XSF (2013)
21. Group, X.X.I.W.: P21451-1-4 standard for a smart transducer interface for sensors, actuators, and devices based on the extensible messaging and presence protocol (XMPP) for networked device communication. Ieee, IEEE Standard Association (2008)
22. Zhang, C., Ding, Y., Ostergaard, J., Bindner, H., Nordentoft, N., Hansen, L., Brath, P., Cajar, P.: A flex-market design for flexibility services through DERs. In: Innovative Smart Grid Technologies Europe (ISGT EUROPE), 2013 4th IEEE/PES. (2013) 1–5
23. Kok, J. K., Warmer, C.J., Kamphuis, I.: Powermatcher: multiagent control in the electricity infrastructure. In: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems, ACM (2005) 75–82
24. Herrmann, I., O'Connell, N., Heller, A., Madsen, H. In: CITIES: Centre for IT-Intelligent Energy Systems in Cities. (2014) 1–8
25. Valsomatzis, E., Hose, K., Pedersen, T. B., Siksnys, L.: Measuring and comparing energy flexibilities. In: Joint EDBT/ICDT PhD workshop. (2015) 78–85

Advanced Museum Services

Maurizio Calderamo[1], Simona Ibba[2], Filippo Eros Pani[2], Francesco Piras[1] and Simone Porru[2]

[1]SoSeBi Srl, Via dell'Artigianato, 9, 09122 Cagliari, Italy
[2]Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi, 09123 Cagliari, Italy
{calderamo, piras}@sosebi.it,
{filippo.pani, simona.ibba, simone.porru}@diee.unica.it

Abstract. The research project aims to develop a prototype of a web-based application specifically designed to improve both the management of and the dialogue between cultural heritage institutions, as well as providing cataloging and web publishing tools for museums, primarily focusing on user interaction. The main goal of the project is to build a product capable of understanding the future evolution of museum services, which can no longer ignore how technological developments are shaping and inspiring users' new habits, together with more advanced and diverse expectations and needs. This project is developed by the Department of Electrics and Electronics Engineering of the University of Cagliari and SoSeBi Srl. It is financially supported by the Autonomous Region of Sardinia with European local development funds.

1 Introduction

Typically, the cultural heritage sector shows an enormous potential for innovation, with astounding development prospectives. This evolution scenario is of particular interest to Italy, being it a country which owns a rich cultural heritage. Such heritage can also be relevant as an economic asset, with fairly interesting and diverse consequences and peculiarities.

Most software offered both in the Italian and international market almost always separate functions related to the description and preservation of artifacts from functions related to the presentation and sharing of the same artifacts with users, without much concern for usability and user experience.

We thus believe that there is more than enough room to introduce an innovative product that exploits the technology potential of web-based architectures.

In a globalized economic context cultural heritage is a strong element of attraction for a region; starting from it, and thanks to ICT, it is possible to trigger disruptive development dynamics.

The use of new technologies can offer a plethora of new opportunities for dissemination and access to cultural heritage, especially to museums. A modern museum can be seen as a complex ICT system, strongly interconnected, and responsible for a vast amount of data to manage. Museums have to be extremely dynamic because of time-changing temporary exhibitions, which are more frequently associated with highly-usable applications to increase visitors' involvement [1], [2].

Aware of the potential shown by ICT when applied to cultural heritage, the proponents aim to start an innovation plan with the objective of developing a prototype of a web-based application dedicated to the management and communication between museums, and to the cataloguing and publication on the web, including the interaction between users. In fact, ICT application to cultural heritage is particularly interesting because of the various innovation opportunities that can be specifically provided to this context.

The latest ICT technologies can trigger a disruptive evolution in a museum system, ranging from information management and communication systems to the creation of a new typology of interactive museum [3].

Thanks to the Internet and to the new means of interaction with information systems (e. g., immersive and natural means), the ways of enjoying a museum are multiplying. In the near future we will be able to think about a cultural space projected towards the external environment not only for pure management purposes, but also for entertaining and educating.

Nowadays, as we can witness the birth of a innovative museum model which is interactive, interconnected, multimedial, and more and more similar to a lively cultural space to the service of a everwider audience, the main goal appears to be that of studying and designing a unique software solution capable of providing the following functionalities:

- traditional cataloguing management for collections hosted in museums;
- management of the exchange of items between museums;
- online publication of the museum collections;
- augmented reality and user interaction (creation of virtual thematic paths with integrated digital information);
- semantic search and automated items cataloguing.

At the present day, museums in Italy are not provided with a single product capable of combining the accuracy requested by cataloguing activities with the user need of accessing museum items on the Internet. From a technological point of view, the software, as for the reasons previously stated, will be installed on a cloud computing web platform which will allow scalable access also for mobile devices.

This paper is structured as follows: the second section describes the prerequisites and motivations behind the proposed project; the third section describes the context of the research proposal. In the fourth section, the schedule of the project is outlined. The last section hosts our final observations about the project.

2 Project Prerequisites and Motivations

A number of national surveys indicate limited growth for museum software applications demand. It is possible to assert that the museum software market has reached the maturity stage. IT companies compete for a market share, economies of scale push for price reduction, and services value is rising. It is precisely to services that the project proponents will address most of their efforts in terms of investments.

Studies and industries surveys predict an increasing demand for new high-technology services, whose valorization will possibly constitute a new source for revenues for all of those companies which will recognize the signs of this growth.

In order to fully satisfy the new needs of the reference market, the new product will be designed and developed exploiting the technological assets that web-based architecture offer, such as the QR code technology, and communication with the main players in the market, including Google. The software will also fulfill all the technical standards promoted by the Cultural Heritage Ministry through the dedicated web portal provided by "Cultura Italia" [1].

The technological innovation of the new product is threefold: quality enhancement, evolved performances, macro-functionalities integration.

The future scenario for cultural heritage in which museums will operate will feature an ever more central role of the web as a means to access knowledge, represented by the collections belonging to cultural heritage institutions. The web itself is facing a constant acceleration in technological innovation, thus offering more and more advanced and sophisticated virtual environment exploration capabilities to users. This makes more room for innovation in all of the contexts, and will deeply modify people lifestyle and interaction on the Internet.

Through virtual tours, cultural tourism will turn into a fundamental development opportunity for public institutions. New ways for accessing cultural heritage, enriched by an immersive user experience for the museum visitors, will be offered to the public. Semantic technology will allow for those software systems capable of effectively using the available assets described in both public and private institutions' repositories to cooperate on the web, at a global level.

The cooperation between these smart agents, together with their interoperability, will integrate all the cultural assets hosted by museums connected through the web, thus making for a more effective and easier to find cultural heritage for the final user, which is currently fragmented due to the vast plethora of technological-divided institutions.

Users will be all the more drawn to share their virtual tours experiences through social networks, thus opening new scenarios for interaction and museum exploration. In fact, those experiences will be shown and suggested to their friends and acquaintances with matching artistic taste. This will in turn create "emotional paths" mediated by their own network of online relationships.

In this scenario, proponents aim at realizing a prototype of a new software for museums, which will have the peculiar feature of integrating five macro-functionalities into a single application. Usually, those functionalities are provided by different software products.

The main goal is to design a product capable of fulfilling the needs originated by the new direction he Cultural Heritage sector is heading to. The application will not be indifferent to the technological developments demanded by the current users needs, whose expectations and demands are more and more advanced; they are also modeled and inspired by consumer technologies promoted by multinationals in the digital sector, such as Google, Apple, Facebook, Amazon. Moreover, the very same users expectations are to be found in entertainment exploration activities in museum environments, which is also an aspect that is gaining momentum because of the need to answer the request for new means of access by younger generations.

[1] Cultura Italia, www.culturaitalia.it

The resulting new generation application, characterized by an innovative view on the concept of museum automation application, is expected to be the pioneer of a general development trend in the museum sector, that is, it will be capable of anticipating and understanding the market evolution in the next years. The current Italian and international market landscape only offers applications which clearly distinguish functionalities that are used for describing and preserving cultural assets from those used for presentation and sharing among users. Less focus is then to be found on usability and user experience, thus making more room in the market for an innovative product featuring the integration of macro-functionalities, when associated with a market proposal as innovative as the product itself.

As a result, the final product will be a network-oriented software, entirely web-based. All the data management processes will be externalized into a Cloud Computing infrastructure, and offered to the users as a service (Software as a Service, or SaaS).

3 The Research Proposal Context

Most of the Italian museums build their peculiar features mainly on their historical background. Taken as a whole, Italian museums originated just after Italy was unified, to avoid losing the cultural heritage as a consequence of the transfer of religious buildings (e. g., churches) ownership to the central state. Therefore, Italian museums were not conceived as depositories for exotic masterpieces intentionally collected, but as places hosting heterogeneous objects, locally collected, and grouped out of necessity.

They were first created in a civic form, and placed inside buildings often of great worth. Italian museums are also more densely aggregated in central and, sometimes, northern regions.

Countless museum typologies exist, classified according to the sector they are concerned with and the objects they host.

There are art museums and historical museums, the latter often dedicated to the city where they are placed or to a main historical period or event, such as the Risorgimento museums or Resistance museums. Always belonging to this typology, there are also the archaeological museums, dedicated to the most ancient objects, and those concerned with a specific civilization, which collect historical and artistic items, such as the Egyptian Museum in Turin, or the various museums dedicated to the Etruscan civilization in Tuscany and Lazio.

In all the cities that count among their past inhabitants a famous person (artist, intellectual, historic character, etc.) it is possible to find a house museum, that is, a museum placed in the very same space where they had worked or lived.

In addition, there are anthropological and ethnographic museums, dedicated to different human civilizations and to their artifacts and endeavours. Besides local history museums, Italy offers many other museums, such as those on rural traditions, traditional craftsmanship (e. g., tailoring and dressmaking, marble and wooden artifacts, pottery), or to a specific food (e. g., chocolate, citrus fruits, olive oil, cheese).

Considering natural science, we must mention the natural history museums, which host collections of animals, plants, minerals, and reconstructions of natural environments; also science and technology museums, documenting the evolution of human discoveries in science.

In Italy, as well as throughout the world, capitals and other major cities are home to the great state museums which arose mainly in the nineteenth century, and that host countless objects and artifacts, mostly from earlier private collections. In recent times, a number of museums specialized on many different topics have sprung up even in small towns, aiming at valorizing and promoting the most peculiar local traditions and finest local products.

There are also museums that have been founded thanks to donations from private collectors. They usually focus on a certain objects typology, such as dolls, toys, stamps, and so on. Recently, in Italy as well as in different parts of the world, a novel museum typology has been established, one specifically designed for children, which offer itineraries and laboratory activities designed for educating children. This kind of museum has become particularly popular among schools and families. Museums and similar institutions are, to a considerable extent, true cultural centers, capable of combining the basic functions of safeguarding cultural heritage, research, and exhibition, with those related to the promotion of educational activities, discussion, information exchange, exhibitions, and contemporary art production, that is, those concerning the entertainment of local communities. In a statistical study published in 2013, ISTAT[2] conducted a survey in collaboration with the Ministry of Heritage and Culture, where regions and autonomous provinces were throughly analyzed to list not only all the Italian museums, but also similar institutions, that is, museum-like institutions, either public or private, either managed by the public administration or not.

4 Research Project Description

The new ICT technologies can radically improve a museum system starting from the information management and communication systems, so as to finally shape a new type of interactive museum. Thanks to the Internet and the new forms of interaction with computer systems now available, users can now choose among different ways of enjoying a collection hosted inside a museum: in the near future, it will be possible to think of an outward-looking cultural area, with more and more recreational and educational services combined with the more usual informational ones.

The expected product will be a modern application with an innovative vision of the concept of automation of museum services, that is, a pioneer software capable of following the current development direction that the museum sector is taking, anticipating and taking advantage of market trends in the years to come.

Our aim is to create a product that can anticipate the direction in which the museum and cultural/artistic heritage sector is evolving. That is a sector that can no longer ignore the technology development fueled by users' new habits, expectations, and needs, which are in turn shaped by consumer technologies coming from multinational companies in the digital industry.

Such expectations interest different aspects, namely game-like exploration of museum space, which is becoming popular to address the learning style of young users.

[2] Istituto Nazionale di Statistica, www.istat.it

The software will be designed for the Internet, as it will be entirely web-based, while the management of all data will be outsourced to a Cloud Computing infrastructure, and will be provided to clients as a service (Software as a Service). The platform will have the distinctive and innovative feature of integrating five macro-functionalities (which are usually managed via different software products) in a single application:

1. traditional management of collections hosted in museums: cataloguing, digitization of museum collections, etc.;
2. management of the exchange of objects between museums: development of an application to explore the entire cultural heritage belonging to each museum;
3. online publication of museum collections: development of a "front-end web portal" for the exhibition of museum collections, which can increase and facilitate the access by a multi-target user base;
4. user interaction and augmented reality: use of QR code to create virtual cultural paths, customized and thematic, according to the new technological paradigm of augmented reality;
5. semantic search: automated items classification in separate categories on the basis of formalized taxonomies or user-generated folksonomies, which will be possible through the interpretation of their descriptive content.

The homogeneous integration of the above functionalities will be an original feature of the software platform. The application will address two different types of users: museums end users and museum curators, which will also be able to easily handle loan requests in order to improve the exchange of items between the various museum institutions.

The prototype will be developed leveraging open source technologies in accordance with the need of the Public Administration for using free software, and, at the same time, to follow recent market trends that testify how FLOSS software is gaining popularity also thanks to the greater social value it offers (especially if compared to proprietary software).

The platform will be integrated with the most widespread social networks to enlarge the targeted customer segment and gather feedback. A new type of visitor is in fact becoming common in museums as well as in most cultural spaces. These visitors want to communicate, and report their experience in a non-conventional way through their smartphones or tablets. Comments and opinions, carried through social channels, will be monitored and analyzed with techniques coming from sentiment analysis and opinion mining, in order to obtain a qualitative confirmation of the appreciation felt towards the museum and its web application by users.

4.1 Project Subdivision

The project officially began on October 1, 2015, and its conclusion is estimated to be on September 28, 2018.

The project covers a number of operation stages. Every stage of the working plan is organized in Work Packages (WP), parallel phases in which operative objectives are reached with a work group activity, through the production of expected results and products and the application of a specific methodology.

The WP included in the projects are five:

WP 1 - Traditional management of museum collections
WP 2 - Management of the exchange of items between museums
WP 3 - Semantic search
WP 4 - Online publication of museum collections
WP 5 - User interaction and augmented reality

Below is a brief description of each phase of the project development.

4.1.1 WP 1 - Traditional Management of Museum Collections

Ensuring compliance with both the current Italian and international industry standards is a fundamental requisite for developing a software capable of dealing with a scientific collection of museum resources. Assuring compliance with standards will allow for a correct and consistent cataloging of items belonging to collections, also according to the legislation of other countries. For this reason the first activity SoSeBi will have to perform is a survey on standards and metadata to apply to the cataloguing and digitization of museum collections in compliance with national and international laws. In addition, with regard to this activity, it will be necessary to perform a survey also on the assets and cataloguing management methodologies for museum collections, again at both national and international level.

Concerning technologies and software development tools used in this project, the proponents intend to undertake, within WP 1, a survey to investigate the possibility of diversifying software production, as currently this process is strictly linked solely to the technological paradigm related to proprietary programming languages. In contrast, the investment through a special consulting activity aims at investigating the possibility of introducing in the production process also Open Source technologies, assessing weaknesses and strengths of open source software development tools and their specific relevance to the realization of the software prototype.

The main proponents goal will be to define the application and the innovative features of the Management Module. On the one hand, the macro-management capabilities developed for collections will be an easy and effective means for the museum operator to manage the descriptions of museum resources and the connections between them. On the other hand, it will offer to the visitor the best and most effective way for enjoying all the museum collections. This macro-functionalities allow for controlling all management activities requested by museums, in a modular way with respect to exchange, publication, interaction, and semantics. To achieve these results, a graphical interface driven by a uniform and consistent operating logic will be employed, thus allowing the back-end operators to make use of the different functionalities without having the feeling of using different software products.

The web interface of the back-end addressed to museum operators provides a customizable dashboard, specifically designed for focusing on the most used functionalities. For example, an operator will have the opportunity to highlight the cataloguing functionalities, whereas the museum director will have easier access to staff management features or statistical data. On the dashboard, all the software functionalities will be easily accessible. Their layout and activation will fully depend on the level and type of the operator and the tasks assigned to them.

Another goal is the creation of a macro-functionality capable of providing the museum curator to aggregate information on the nature, extent, and structure of the

assets hosted by the museum. The application will allow for the management of the information associated to each resource according to a paradigm based on the concept of complex networks, integrating “atomic” information (intended as referring to the individual item, thus according to traditional cataloguing) with another type of information associated with the connections between the various resources constituting the assets of the museum (gallery, exhibition, etc.). This second type of information is of a "topological" nature [4].

The application will be capable of generating, from the information contained in the catalog, a complex network where the assets and their reciprocal relationships will be made clearly observable (in aggregate form and at different levels of granularity). This presentation will also be visualizable in a schematic form easily and immediately explorable, and will investigate the complexity of the existing relationships between items.

The proposed network is, in fact, a mathematical model on which to apply various algorithms for network metrics detection, and perform analysis of a statistical nature. For instance, it will be possible to obtain information on the nature and strength of the connections between the assets through the use of clustering algorithms, or community detection algorithms. The temporal evolution of the hosted resources will be observable, and it will be possible to compare information on the resources with information of other nature (e. g., geographical information). It will also be possible to filter network elements on the basis of different parameters (period of history, style, author, etc.), and then perform analysis only on subnets of interesting elements.

4.1.2 WP 2 - Management of the Exchange of Items between Museums

Another objective to achieve will be the creation of an application macro-functionality allowing the operator to optimize the exchange of objects between museums, thus making it possible to easily organize thematic cultural events.

The application will suggest optimal time and date for the scheduling of thematic exhibitions on the basis of predefined constraints, such as the availability of the items pertaining to the chosen theme. Each work will be eventually associated to a cost comprising booking, transportation, and management. Date and place will be chosen on the basis of the solution to an optimization problem, whose performance index represents the potential profit deriving from the organization of the event.

The macro-functionality described within WP1 might find its best application in a scenario where access to museum objects were extended towards other entities than the museum, that is, for example, external users. In this context, the application will be designed so to be able to integrate topological information on the reference structure with those from other museums.

Starting from the information extracted from the local catalogue and those retrieved from the catalogues of other museums, the application will also be able to generate a complex network where all the items and the relationships between them will be clearly distinguishable (in aggregate form and at different levels of granularity) [5].

The curator will be able to assess the nature of the relationships between the artifacts kept in the institution where he works and those hosted in other institutions, in order to evaluate any opportunities for partnerships in the organization of events, of guided tours, or in drafting catalogues [6].

4.1.3 WP 3 - Semantic Search

This WP concerns the realization of a module whose objective is to automatically (or semi-automatically) generate a taxonomy starting from the available textual data on the museum resources.

The project will feature automated classification of catalogued assets in specific categories on the basis of taxonomies built through a detailed analysis over the textual description of the resources. This objective will be achieved through the correct interpretation of the related descriptive content.

Taxonomies are hierarchical structures aimed at organizing information; their use allows to improve the performances of information retrieval systems, such as vertical search engines. Vertical search engines typically use ad-hoc taxonomies which describe the domain of interest. This happens in cases where it is needed to provide a specific service (e. g., a service addressed to tourism) and/or a localized service (e. g., addressed to tourism within a specific region). Currently, the problem of the generation of ad-hoc taxonomies can be dealt with by using groups of experts which operate without the aid of automated tools [7], [8]. In addition to the long realization time this approach requires, it is also often difficult to assess the adequacy of the generated taxonomies, with a negative impact on the performances of the search engine which is using the taxonomies. It must be also considered that, as of today, the application domains are changing very quickly. Thus, employing a manual approach is not suitable for promptly updating the taxonomy. For these reasons, the study and the definition of automated (or semi-automated) solutions for taxonomy generation, specifically tailored to the domain of interest, is steadily increasing, due to the plethora of applications in which they could be exploited.

WP3 also aims to improve the relevance of the results given as a response to user queries. The most suitable way for effectively and efficiently finding relevant information on museum artifacts both for visitors and curators is using a vertical search engine, that is, a search engine focused on a particular area, able to get to the most detailed information about indexed items (in this case, museum artifacts) and quickly and effectively return the most relevant results to the user.

The main goal of the WP is then to define and develop suitable techniques and algorithms for automated or semi-automated taxonomy building, specifically tailored to the cultural heritage domain, and subsequentely implement a vertical search engine on top of it.

4.1.4 WP 4 - Online Publication of Museum Collections

One of the main objectives within WP4 is the definition of the application features and the innovative functionalities of the Publishing Module. This objective is related to the definition and detailed description of the application functionalities, also with regard to the information exchange with the other application modules and the subsequent production of the document for the analysis that will reveal which information are necessary to identify the software development guidelines for the developers team.

Another objective within WP4 will be to pay attention, during the online publication phase, to the web interface provided by the application. It will be designed with a user-centric approach, to establish a connection between the observer and the observed object to make the individual the main protagonist in the interaction.

The web front-end implementation, which will follow an immersive logic, will present museum exhibitions as a open spaces, seamlessly integrated with the context of cultural heritage tourism, and asking for continuous user interaction.

Therefore, context, space, and time, acquire great relevance: the experiential interface is the place where relationships between individuals, places, historical periods, businesses, and museum artifacts, are established.

The visitor is not a mere observer, as he interprets the object through an interaction that starts online, before he visits the museum: the approach toward the museum object is properly customized according to the visitor personal profile. Profiling allows for customized browsing of that content considered as the most attracting to the visitor, according to the user typology (e. g., school, critic, historian).

Through visual elements such as the front-end layout, multimedia, flash animations, and 3D, the user is fully involved and invited to experience the effects of his presence and actions when interacting with a museum exhibit or a single object. The object acquires more value as the frequency of the interactions with users grow. Digital information enhances the fundamental principles of any museum exhibition: socialization as shaped by the museum communication strategy; the full understanding of the real world which lays the foundations for a sensorial, practical approach to the observation of the exhibits, as opposed to a virtual approach; the vision of humanity as a multi-sensory experience able to enhance the human ability to relate to both the real and the virtual environment. The user physical presence is at the center of his online visit. The narrative associated to the various museum exhibits is thus made explicit through a series of both original and user-generated multimedia content (which are also highly integrated with the social platforms). The user experience generates strong expectations about the exhibit: the object is presented to the visitor not only as something esthetically enjoyable, but as an entity not separable from the context it belongs to, with a special meaning.

The integration of artifacts, historical content, and user-generated content is a innovative feature of the platform. The realization of the experiential interface requires leveraging a deepest know-how on various fields, even profoundly different ones. The achievement of the general and specific objectives of the project is possible only through cooperation of the best actors belonging to different sectors.

4.1.5 WP 5 - User Interaction and Augmented Reality

Another key objective will be the definition of the application features and the innovative functionalities of the interaction module. In addition, DIEE and SoSeBi will work together on the creation of a macro-functionality whose aim is enabling the visiting user to optimize their expendable time in a city of art or a large museum on the basis of their cultural interests and context constraints [9], [10].

First and foremost, each user will be able to include in their profile their preferences in terms of authors (painters, sculptors, architects, etc.) and will be constantly updated on exhibitions and potentially interesting events.

Second, each user will be allowed to ask for a detailed tour plan about visits to cities of art or great museums on the basis of a series of information the user provides to enrich their profile. For example, user will be able to maximize the time spent in a city, not only on the basis of the priorities he expressed, but also on constraints such as opening hours and distance. Similarly, the user can optimize their visits to museums on

the basis of the available time, their preferences, recommendations from other users, information from their previous visits, information on temporary exhibits, etc.

Another goal that will be achieved will be the creation of a social network composed by the system users wishing to exchange and share information with other users showing similar tastes and interests.

It will be possible to enhance the service quality of the user decision support system taking advantage from the social interaction between users. The users will be invited to establish relationships with other users properly selected by the system on the basis of the similarity ratio between profiles and experiences. This will be useful both to users who access the service as museum (or, more generally, cultural site) visitors, and users who access it as curators of events, exhibitions, or sites of various typologies.

The social dimension of the system allows for establishing direct relationships between new users and experienced users (i. e., those who have already enjoyed a given experience in a city) thus showing the latter's opinion on the visit and, eventually, receive immediate feedback and suggestions to improve their own experience.

The analysis of the interactions between the users and the information offered by social networks will be useful to the events/sites curators, as it will serve as a tool both for reviewing and validating their work, and for improving what future events will have to offer, social marketing, and expected data about tourists participation to a given event, on the basis of the activity recorded on social networks and the number of scheduled tours planned by the users.

The platform to be realized will be integrated with the most popular social networks in order to expand the reference customer segment and collect as much users feedback as possible. In fact, for museums, as well as for most of the cultural points of interest, a new visitor typology is starting to acquire relevance, that is, the visitor who wants to communicate and that, with the help of handheld devices such as smartphones or tablets, wants to relate their experiences in an innovative way. The comments and opinions conveyed through social media will be monitored and analyzed through the use of sentiment analysis and opinion mining techniques. This will allow for obtaining feedback about the perceived quality of a specific museum, together to that of the web applications it provides.

4.1.6 Protection and Exploitation of the Results

Protection and exploitation of the project results will be ensured via the following actions:

1. deployment of the new product, as the summary of the results obtained from basic and industrial research activities, and subsequent commercialization. It must be taken into account that the product is the tangible form of its intellectual content. As exploitable at different levels of research and engineering, its market penetration will be, especially from the company point of view, the most effective valorization effort for the content it represents;
2. generation of a momentum effect for the creation of even new experiences and acquired know-how through the activation of a networking mechanism, that is, the creation of a system built upon the relationships with research institutions within the region and the country (especially, with the University of Cagliari through the DIEE, and also with other companies operating in the ICT and cultural heritage sector);

3. production of scientific publications accounting for the results of the research effort performed by the proponents.

5 Conclusion

The future of cultural heritage, in which museums will be called to act, will see the web as an increasingly central means of access to knowledge and, more specifically, to artifacts hosted by museums and other cultural institutions. Even the web is experiencing a continuous multiplication of technology innovation processes, thus offering advanced and sophisticated opportunities for exploring virtual environments and for education.

To fulfill the new needs of the reference market (i.e., the customer segment addressed by public and private museums) we chose to design and develop a new generation software product, with an innovative vision on the very concept of museum automation program. This product could pave the way to a disruptive development trend in the cultural heritage sector.

The presented project is financed by the Autonomous Region of Sardinia with European funds (Single Programming Document 2007-2013 - P.O. FESR 2007-2013 – Line of Activity 6.2.2.d - Interventions to support competitiveness and innovation, under the Regional Committee Resolution no. 33/41 of 08/08/2013).

The creation of the platform will stem from the strategic partnership between SoSeBi Srl and the Department of Electrical and Electronic Engineering (DIEE) of the University of Cagliari. The motivation behind this choice is to use the results obtained both from fundamental research and industrial research to elaborate an innovative prototype, that aims at being unique in the domestic market as for innovative features it will offer.

This project is coherent with the strategic objective of the regional planning in Sardinia, since it aims to implement innovative methods of the ICT sector in the library industry, and it complies with the objectives described in the Regional Strategic Document (Documento Strategico Regionale, DSR) 2007-2013 for Sardinia.

Acknowledgements. Simona Ibba and Simone Porru gratefully acknowledges Sardinia Regional Government for the financial support of his PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 Axis IV Human Resources, Objective l.3, Line of Activity l.3.1).

References

1. Bartolini, I., Moscato, V., Pensa, R. G., Penta, A., Picariello, A., Sansone, C., Sapino, M. L.: Recommending multimedia visiting paths in cultural heritage applications. In: Multimedia Tools and Applications (2014), pp. 1-30, DOI: 10.1007/s11042-014-2062-7.
2. Bitgood, S.: An analysis of visitor circulation: Movement patterns and the general value principle. In: The Museum Journal (2006), Vol. 49, Issue 4, pp. 463-475.

3. Chianese, A., Piccialli, F.: Designing a smart museum: when Cultural Heritage joins IoT. In: Eighth International Conference on Next Generation Mobile Apps, Services and Technologies (NGMAST), IEEE (2014), pp. 300-306.
4. Newman, M. E. J.: The structure and function of complex networks. In: SIAM REVIEW (2003), 45:167-256.
5. Clauset, A., Moore, C., Newman, M. E. J.: Structural Inference of Hierarchies in Networks. In: Airoldi, E. M. et al. (Eds.), ICML, Lecture Notes in Computer Science 4503, Springer-Verlag, Berlin Heidelberg (2007), pp. 1-13.
6. Clauset, A., Moore, C., Newman, M.E.J.: Hierarchical structure and the prediction of missing links in networks. In: Nature 453 (2008), pp. 98-101.
7. Skinner, J.: Metadata in Archival and Cultural Heritage Settings: A Review of the Literature. In: Journal of Library Metadata (2014), Vol. 14, Issue 1, pp. 52-68. DOI: 10.1080/19386389.2014.891892.
8. Pani, F. E., Concas, G., Porru, S.: An Approach to Multimedia Content Management. In: Proceedings of the 6th International Conference on Knowledge Engineering and Ontology Development, KEOD (2014), pp. 264-271.
9. Kelly, L.: The Connected Museum in the World of Social Media. In: Museum Communication and Social Media: The Connected Museum, New York-London: Routledge (2014), pp. 54-71.
10. Rubino, I., Xhembulla, J., Martina, A., Bottino, A., Malnati, G.: MusA: using indoor positioning and navigation to enhance cultural experiences in a museum. In: SENSORS (2013), Vol. 13, Issue 12, pp. 17445-17471. ISSN: 1424-8220.

PERICLES – Digital Preservation through Management of Change in Evolving Ecosystems

Simon Waddington[1], Mark Hedges[1], Marina Riga[2], Panagiotis Mitzias[2], Efstratios Kontopoulos[2], Ioannis Kompatsiaris[2], Jean-Yves Vion-Dury[3], Nikolaos Lagos[3], Sándor Darányi[4], Fabio Corubolo[5], Christian Muller[6] and John McNeill[7]

[1]King's College London, U.K.
[2]Information Technologies Institute, CERTH, GR-57001 Thessaloniki, Greece
[3]Xerox Research Centre Europe (XRCE), 38240 Meylan, France
[4]Swedish School of Library and Information Science, University of Borås, Borås, Sweden
[5]IPHS, University of Liverpool, L69 3GL, U.K.
[6]B.USOC, Brussels, Belgium
[7]Tate, London, U.K.
{simon.waddington, mark.hedges}@kcl.ac.uk,
{mriga, pmitzias, skontopo, ikom}@iti.gr,
{Jean-Yves.Vion-Dury, Nikolaos.Lagos}@xrce.xerox.com,
Sándor.Darányi@hb.se, corubolo@gmail.com,
christian.muller@busoc.be, john.mcneill@tate.org.uk

Abstract. Management of change is essential to ensure the long-term reusability of digital assets. Change can be brought about in many ways, including through technological, user community and policy factors. Motivated by case studies in space science and time-based media, we consider the impact of change on complex digital objects comprising multiple interdependent entities, such as files, software and documentation. Our approach is based on modelling of digital ecosystems, in which abstract representations are used to assess risks to sustainability and support tasks such as appraisal. The paper is based on work of the EU FP7 PERICLES project on digital preservation, and presents some general concepts as well as a description of selected research areas under investigation by the project.

1 Introduction

1.1 Motivation

Existing approaches to digital preservation are heavily influenced by practices that have evolved over many years in the non-digital world. The reusability of digital objects is dependent on their surrounding environment. This can include not only relevant software, but also platforms and documentation, and often the digital objects and their environment have complex interdependencies. Due to the rapid pace of technological change, the environment in which a digital object exists will evolve and some entities may become delinked or even obsolete. This may result in the loss of capabil-

ity to run software, to interpret information or to render data files. A similar argument can be applied to other types of change. For example, organisational changes may result in digital objects held by the organisation no longer being compliant with current policies and procedures. Evolution of user communities may result in the digital objects being interpreted by individuals and used for purposes that were not envisaged when they were initially created or acquired. This can result in the digital objects not being fit for purpose or even understandable by current users.

Maintaining digital objects in a static form in a repository, as might be done with non-digital assets such as books and paintings, which can remain in a reusable form for centuries, is unlikely to be successful even over time periods of a few years. Thus new approaches are required to managing such digital objects that can deal with both complex dependencies as well as continual change.

1.2 PERICLES Objectives and Approach

The main challenge for PERICLES is to ensure the ongoing interpretation and reusability of digital objects that are heterogeneous, volatile (i.e. subject to continual change) and are complex (i.e. have many interdependencies). By analogy with biological systems, we use the term *digital ecosystem* to reflect an evolving set of interdependent entities, which is subject to influences bringing about change. Digital ecosystems can include any entities that can have a direct or indirect impact on the reuse of digital objects, including data objects, software, user communities, processes, technical services and policies. An important feature of our approach is that a definition of a digital ecosystem includes descriptions of the dependencies between the constituent entities.

Following a widely adopted methodology in science, we introduce computational models to enable the impact of change on a digital ecosystem to be assessed, and in some cases for mitigating actions to be determined, without the need to manipulate the entities in the ecosystem directly. Based on a linked data paradigm, the models make use where possible of existing domain ontologies. The evolution of the models is governed by policies.

In order to populate such models, tools are provided to support the extraction of metadata, such as content, environmental, usage and provenance information. Analysis and visualisation of the models is used to support risk analysis and provide decision support. Finally preservation actions can be determined and translated into executable business processes.

To support the development, testing and real-world deployment of PERICLES components, an integration framework is under development. This includes an Entity Registry and Model Repository to support the storage and retrieval of the models as well as an execution layer to enable preservation components to be wrapped in handlers and run against the stored entities. The integration framework also provides a reference implementation for deploying PERICLES components in real-world applications.

1.3 Digital Preservation Activities in the EU

In this section, we briefly review prior EU-funded activities relating to digital preservation, to place PERICLES in a wider context. In 2001, the EU funded the Electronic Resource Preservation and Access Network (ERPANET) project[1] in the FP5 programme. This was the first attempt to engage with memory organisations, such as museums and libraries, and commercial sector for the purpose of raising awareness about the need for digital preservation and at the same time providing the necessary knowledge base to all participants.

There then followed approximately €100M of EU funding for digital preservation, covering a wide range of topics and including research and development prototypes. The PLANETS[2] project developed the Planets Suite, comprising a preservation planning tool, a test-bed and an interoperability framework. The planning tool, Plato, offered information on digital objects at risk, and supported informed decision-making on preservation actions. The project primarily dealt with simple digital objects, rather than complex dependencies, and used a sampling approach on individual objects to evaluate preservation actions.

CASPAR[3] worked primarily on preservation approaches to validate the OAIS reference model [13] in the cultural, artistic and scientific domains. It investigated the implementation and use of key OAIS concepts such as representation information, knowledge management and preservation description information. SHAMAN[4] studied the incorporation of Product Lifecycle Management within a digital preservation system, and produced an extended information lifecycle model. PROTAGE[5] explored the use of software agents targeting automation of digital preservation processes. The LiWA[6] project dealt with archiving of web content.

TIMBUS[7], addressed preservation of business processes where software and platform are developed and delivered as a service. SCAPE[8] focused on scalable preservation algorithms, extending the results of PLANETS to high volume content. The ENSURE[9] project considered scalable pay-as-you-go infrastructure for preservation services based on cloud computing technology, as well as exploring non-traditional domains for digital preservation such as finance and medicine. Finally the APARSEN[10] network tried to join together work on pervious preservation projects into a common vision, again underpinned by the OAIS model.

PERICLES differs from most preceding projects in that it considers continuously changing environments such as for time-based media, where OAIS is less appropriate. Our approach is based on a continuum viewpoint. Although static dependency models

[1] http://www.erpanet.org/index.php.
[2] http://www.planets-project.eu/
[3] http://www.planets-project.eu/
[4] http://shaman-ip.eu/
[5] http://www.ra.ee/protage
[6] http://liwa-project.eu/
[7] http://timbusproject.net/
[8] http://www.scape-project.eu/
[9] http://ensure-fp7-plone.fe.up.pt/site
[10] http://www.alliancepermanentaccess.org/index.php/aparsen/

were considered in earlier projects, such as CASPAR, the use of dynamic models is new. There has also been relatively little work done to date on semantic change, which is an important focus of PERICLES.

1.4 Acknowledgements

This work was supported by the European Commission Seventh Framework Programme under Grant Agreement Number FP7-601138 PERICLES. The authors wish to acknowledge the contributions of our many PERICLES colleagues to the content of this paper.

2 Digital Preservation and Change

2.1 Lifecycle versus Continuum Approaches to Digital Preservation

Lifecycle models are a point of reference for most existing approaches and practices in digital preservation. They provide a framework for describing a sequence of actions or phases, such as creation, productive use, modification and disposal, for the management of digital objects throughout their existence. Such models suggest a linear sequence of distinct phases and activities, which in practice may be non-linear or even relatively disordered. Lifecycle models provide an idealised abstraction of reality, and may typically be used in higher-level organisational planning and for detecting gaps in procedures.

The DCC lifecycle model [10] is one of the most well-known lifecycle models. It provides a graphical, high-level overview of the stages required for successful curation and preservation of data from initial conceptualisation or receipt through the iterative curation cycle. The UK Data Archive describes a research data lifecycle[11], which comprises six sequential activities and, unlike the DCC model, is more focused on the data user's perspective. Overviews of lifecycle models for research data are provided by Ball [11] and the CEOS Working Group on Data Life Cycle Models and Concepts [12].

So-called lifecycle approaches typically envisage a clear distinction between active life and post-active life. The Open Archival Information System (OAIS) [3] is a commonly adopted reference model for an archive, consisting of an organisation of people and systems that has accepted the responsibility to preserve information and make it available for a designated community. Although lifecycle models and OAIS provide a useful frame of reference for preservation, they are less suited to dealing with examples where there is a less clear distinction between the active life and archival phases, examples of which will be discussed in section 3.

In [2], we introduced a continuum approach to digital preservation that combines two main aspects. Firstly, there is no distinction made between active life and post-active life; that is, preservation is fully integrated into the active life of the digital

[11] http://ijdc.net/index.php/ijdc/article/view/69

objects. A second aspect is that preservation is non-custodial, that is we do not aim necessarily to remove entities from their environment, both physical and organisational, and place them in the custody of a third party.

Continuum approaches have been proposed in the closely related field of record keeping. The Records Continuum (RC) was originally proposed by Upward in 1996 [4]. An essential aspect is that the content and structure of a record are fixed, but the surrounding context can change over time, so a record is "always in a state of becoming" [15].

2.2 Change Types and Their Impact

PERICLES considers a number of different types of change that can potentially have an impact on the reuse of digital objects. A more extensive review is presented in [16]. The main high-level change types that we have considered are summarised in Table 1.

Table 1. Types of change occurring on digital ecosystems and their impact.

Type of change	Description	Impact
Knowledge and terminology	Changes in semantics that originate from a designated user community.	Different user communities using the same underlying datasets with different understanding and goals.
Technology	This includes hardware availability, software obsolescence, and changes in formats, protocols and interfaces.	Requires replacement of hardware and software components, transcoding of files, redesign of interfaces etc.
Policy	Changes in permissions, legal requirements, quality assurance and strategy.	This can impact how and where digital objects are stored, quality processes they are subjected to, retention periods etc.
Organisation	Change to the organisation due for example to political, financial or strategic reasons. Often organisational changes can be manifested as policy changes.	This can result in different priorities for retaining or maintaining the reusability of digital objects.
Practice	This change originates from new or changed habits of the designated user community (not necessary related to knowledge and terminology changes). It is an indicator that user requirements may change.	This can result in changes to the form in which digital objects are retained, reflecting the changing ways in which they are to be reused.
Requirements	This can include business requirements, functional requirements that a system should fulfil, quality of service and user requirements.	This again reflects the way that digital objects are reused and hence how they should be stored and maintained.
Dependency	Either characteristic attributes of a dependency are changed (e.g. quicker, faster, more flexible, cheaper) or the dependency itself changes.	Evolution in dependencies can reflect different views on the types of change that are being considered.

2.3 Change and Dependency

Change and dependency can in many respects be viewed as dual notions. Thus the types of dependency we may wish to model are related to the types of change that are being addressed. In PERICLES, we say that entity A is dependent on entity B if changes to B have a significant impact on the state of A. A key aspect of PERICLES is that dependencies can have associated semantics and do not merely represent a link between the two objects. The semantics of a dependency are related to the change context under consideration.

A number of notions of dependency exist in the literature. The PREMIS Data Dictionary[12] defines three types of relationships between objects: structural, derivation and dependency. In particular, a derivation relationship results from the replication or transformation of an object. A dependency relationship exists when one object requires another to support its function, delivery, or coherence.

The *Open Provenance Model* (*OPM*)[13] introduces the concept of a provenance graph that aims to capture the causal dependencies between entities. The most relevant concept from our perspective is *process* that represents actions performed on or caused by artefacts, and resulting in new artefacts.

In a preservation context, [17] defines notions of module, dependency and profile to model use by a community of users. A *module* is defined to be a software/hardware component or knowledge base that is to be preserved, and a profile is the set of modules that are assumed to be known to the users. A *dependency relation* is then defined by the statement that module A depends on module B if A cannot function without B. For example, a README.txt file depends on the availability of a text editor (e.g. Notepad). The authors of [8] also define the more specific notion of task-based dependency, expressed as Datalog rules and facts. In [19], the notion of task is extended to *intelligibility*, which allows for typing dependencies. The PERICLES modelling approach goes one step further toward genericity, by allowing any kind of dependency specialisation, and provides a much richer topology for dependency graphs through managing dependencies as objects instead of properties.

2.4 Semantic Change

An important aspect of PERICLES is the study of *evolving semantics* and *semantic change* in particular. The risk of semantic change for digital preservation is that, as a fallout from inevitable language evolution that has been accelerating due to an interplay of factors, future users may lose access to content, either (a) because the concepts and/or the words as their labels will have changed, or (b) because the same concept may have different labels over separate user communities.

Additionally, by better understanding semantic change processes, one can identify ‘at risk’ terminology, which is likely to be hard to understand by future users of the resources. Similarly, one can identify specialist terminology likely to be different across domains and, therefore, difficult for those other domain users to understand.

[12] http://www.loc.gov/standards/premis/

[13] http://eprints.soton.ac.uk/271449/ 1/opm.pdf

Another issue for semantic change might relate to the change over time in the way in which a resource is used.

In the investigation of semantic change, PERICLES is conducting experiments in two major directions: one looking at the interpretability of digital objects over time and over designated user communities and the other focusing on drift[14] detection, measurement and quantification methods. The aim is to eventually relate the two in a common frame of thought.

When considering *drift interpretation* (understandability), one of the key questions is how to represent the knowledge model of a domain, community, or individual. This is a critical question as any subsequent analysis or experimentation will depend upon the quality and 'soundness' of any methods or assumptions made at this initial stage. Regarding *drift quantification* (measurability), vector- versus graph-based measures are computed and ranked. In this way, all kinds of shifts could be analysed, be they community-dependent or temporal.

A key element of our explorations is to address drifts in *word meaning* used for document indexing, and consequent changes in *document meaning* together with *topic shifts typical of evolving document collections*. To this end, we also call in probabilistic methods such as additive regularisation based topic models [20]. Secondly, as in classical mechanics, physical systems in change are typically analysed by *calculus* and represented as *vector fields*. Using physics as a metaphor we introduced a vector field based tool to study evolving semantics [21] and its scalability aspects [22]. This work is in the phase of adding qualitative evaluation of shifts in word meaning to the model, which is novel because typically, only quantitative drifts have been addressed by measurements [24]. Finally, our vector field model of semantic change points in the direction of *social mechanics* [25, 26], thereby paving the way for an integrative meta-theory of changes in sign systems as a function of social use depending on evolving sign contexts.

3 Case Studies

The examples selected for study in PERICLES are chosen from the application domains that the project is addressing, namely digital media, and space science.

3.1 Examples from the Digital Media Domain

Within the PERICLES project, the digital media domain covers three different subdomains, namely Digital Video Art (DVA), Software-Based Art (SBA) and Born-Digital Archives (BDA). Several key challenges have been defined within each of these subdomains and corresponding ontologies have been developed; these do not attempt to model the respective subdomains exhaustively, but are primarily aimed at modelling preservation-related risks. Specifically, in DVA, the focus is on the consistent playback of digital video files, with respect to the technical or conceptual char-

[14] In literature, semantic change is covered by expressions like *semantic drift*, *semantic shift*, *semantic decay*, and sometimes as *concept drift*.

acteristics of the corresponding digital components. In SBA, the focus is on the assessment of risks for newly acquired artworks, regarding their technical dependencies but also the functional, conceptual and aesthetic intentions related to the significant properties of an artwork. Finally, within the BDA context, the focus is on the need to be able to access and maintain digital documents as was originally intended, together with all their technical, aesthetic and permission characteristics. A detailed analysis of the media case study is contained in [27].

All key challenges are explored in relation to collections held by Tate[15]. The DVA and SBA collections belong to the main art collection, while the BDA material exists in the Tate Library and Archive collection; these two types of collections are managed by different teams within Tate. The institution has approximately 300 video artworks, including digital video artworks, in the collection. It also has a small but growing number of software-based artworks. The born-digital material in Tate Library and Archive includes material from institutions within the UK, such as records from commercial galleries that come into the archive, as well as artists' personal records. Much of this material comprises standard formats such as emails, spreadsheets, text documents, images, and so on.

3.2 Examples from the Science Domain

B.USOC[16] supports experiments on the International Space Station (ISS) and is the curator of both collected data and operation history. B.USOC chose to analyse the SOLAR payload, in operation since 2008 on the ESA COLUMBUS module of the ISS. These observation data are prime candidates for long-term data preservation, as variabilities of the solar spectral irradiance have an influence on Earth's climate, and the measurements cannot be repeated. The current SOLAR module is built from three complementary space science instruments (see Fig. 1) that measure the solar spectral irradiance with an unprecedented accuracy.

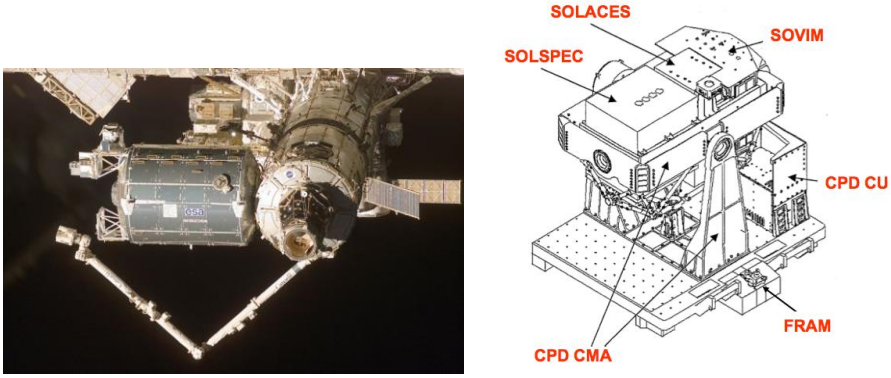

Fig. 1. The International Space Station, and the SOLAR module as part of the COLUMBUS (left) and the SOLAR instrument (right).

[15] http://www.tate.org.uk/
[16] http://www.busoc.be

SOLAR and the phenomena it studies are an example of change in observation data with the evolution of knowledge on the sun. Before the space age (forty years ago), the sun's input to the earth system was called the "solar constant" and great pains were taken to determine it by removing the atmospheric effects. Then in the beginning of the 1980's, several instruments, including first versions of SOLAR, were flown in space and determined that the "constant" was in fact a variable parameter synchronous with the 11 year sunspot cycle, hence its name was changed to "total solar irradiance". Moreover, spectral variations were found not be uniform and that the ultraviolet region, while weak in energy, had important variations relating to solar activity known now as space weather (solar flares and other phenomena).

The SOLAR instruments, which had been designed originally to provide snapshots of the solar spectral irradiance at well-defined parts of the solar cycle, now deliver valuable scientific data relevant to both shorter and longer time scales.

From a digital preservation perspective, the experiments consist of highly complex interlinked digital entities, including raw data and associated telemetry, software, documentation (over a hundred document categories), and operational logs.

4 Model-driven Approach

4.1 Functional Architecture

Following a common paradigm in science, we introduce models to enable the impact of change on a digital ecosystem to be assessed, and in some cases for mitigating actions to be determined. This principle is illustrated in Fig.2, which describes the PERICLES functional architecture, based on [5].

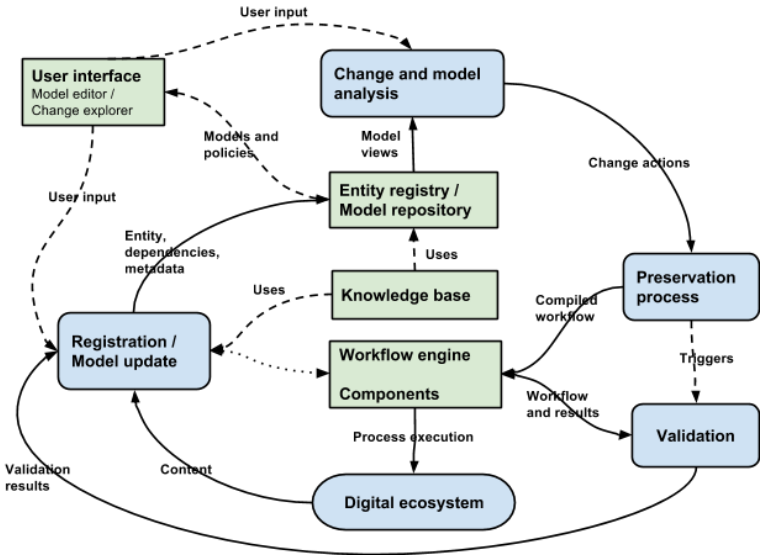

Fig. 2. The PERICLES functional architecture.

The components in blue, joined by solid lines, represent the main workflow for building the models, performing change impact analysis, determining preservation processes or actions, validating the results of the preservation actions, and updating the models. The remaining components support the model-driven workflow. A user interface component is required both to create and in some cases populate the models, as well as to perform change impact analysis and determination of preservation actions. The Entity Registry/Model Repository (ERMR) assigns unique identifiers to the entities in an ecosystem and provides tools for storing and retrieving the models themselves. The Knowledge Base provides the underlying ontologies for constructing the ecosystem models, to be described in the following section, together with reasoning tools. Finally the preservation actions are executed via a workflow engine, using a set of transformation components.

4.2 Digital Ecosystem Models

The PERICLES *Linked Resource Model* (*LRM*) is an *upper level ontology* designed to provide a principled way to modelling evolving ecosystems, focusing on aspects related to the changes taking place. This means that, in addition to existing preservation models that aim to capture provenance and preservation actions, the LRM also aims at modelling how potential changes to the ecosystem, and their impact, can be captured. It is important to note here that we assume that a policy governs at all times the dynamic aspects related to changes (e.g. conditions required for a change to happen and/or impact of changes). As a consequence, the properties of the LRM are dependent on the policy being applied. At its core the LRM defines the ecosystem by means of constituent entities and dependencies. The main concepts of the static LRM are illustrated in Fig. 3. (The prefix `pk` refers to the LRM namespace).

Resource. Represents any physical, digital, conceptual, or other kind of entity; entities may be real or imaginary and in general comprises all things in the universe of discourse of the LRM Model. A resource can be *Abstract* (c.f. `AbstractResource` in Fig. 3), representing the abstract part of a resource, for instance the idea or concept of an artwork, or *Concrete* (c.f. `ConcreteResource`), representing the part of an entity that has a physical extension and is therefore characterized by a location attribute (specifying some spatial information). These two concepts can be used together to describe a resource; for example, both the very idea of an artwork, as referred by papers talking about the artist's intention behind the created object, and the corresponding video stream that one can load and play in order to manifest and perceive the artwork. To achieve that, the abstract and concrete resources can be related through a specific `realizedAs` predicate, which in the above example could be used to express that the video file is a concrete realization of the abstract art piece.

Dependency. An LRM `Dependency` describes the context under which change in one or more entities has an impact on other entities of the ecosystem. The description of a dependency minimally includes the intent or purpose related to the corresponding usage of the involved entities. From a functional perspective, dedicated policies/rules further refine the context (e.g. conditions, time constraints, impact) under which change is to be interpreted for a given type of dependency.

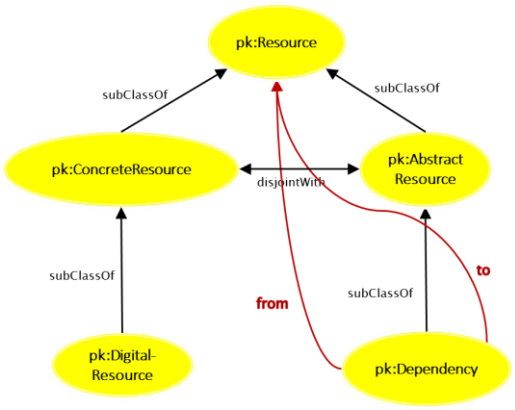

Fig. 3. Main concepts of the static LRM.

For example, consider a document containing a set of diagrams that has been created using MS Visio 2000, and that a corresponding policy defines that MS Visio drawings should be periodically backed up as JPEG objects by the work group who created the set of diagrams in the first place[17]. According to the policy, the work group who created the set of JPEG objects should be able to access but not edit the corresponding objects. The classes and properties related to `Dependency` can be used to describe each such conversion in terms of its temporal information and the entities it involves along with their roles in the relationship (i.e. person making the conversion and object being converted), as other existing models. In addition, the LRM `Dependency` is strictly connected to the intention underlying a specific change. In the case described here the intent may be described as "*The work group who created the set of diagrams wants to be able to access* (*but not edit*) *the diagrams created using MS Visio 2000. Therefore, the work group has decided to convert these diagrams to JPEG format*" and it implies the following.

- There is an explicit dependency between the MS Visio and JPEG objects. More specifically, the JPEG objects are depending on the MS Visio ones. This means that if an MS Visio object 'MS1' is converted to a JPEG object, 'JPEG1', and 'MS1' is edited, then 'JPEG1' should either be updated accordingly or another JPEG object 'JPEG2' should be generated and 'JPEG1' optionally deleted (the description is not explicit enough here to decide which of the two actions should be performed). This dependency would be especially useful in a scenario where MS Visio keeps on being used for some time in parallel to the JPEG entities being used as back up.
- The dependency between 'MS1' and 'JPEG1' is unidirectional. Actually, JPEG objects are not allowed to be edited and, if they are, no change to the corresponding MS Visio objects should apply.
- The dependency applies to the specific work group, which means that if a person from another work group modifies one of the MS Visio objects, no specific conver-

[17] This example is adapted from a use case described in [19], pp. 52-53.

sion action has to be taken (the action should be defined by the corresponding policy).

To enable recording the intent of a dependency, we can relate in the LRM the `Dependency` entity with an entity that describes the intent via a property that we name *intention*, as illustrated in Fig. 4.

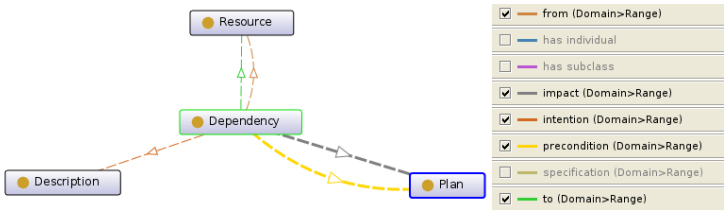

Fig. 4. A view of the `Dependency` concept in LRM.

Let us take once more the example above: we need to be able to express the fact that a transformation to the JPEG is possible only if the corresponding MS Visio object exists and if the human that triggers the conversion has the required permissions to do that (i.e. belongs to the specific workgroup). The impact of the conversion (generating a new JPEG object) could also be conditioned on the existence of a corresponding JPEG object containing an older version of the MS Visio object. The actions to be undertaken in that case, would be decided based on the policy governing the specific operation. Assuming that only the most recent JPEG object must be archived, the old one must be deleted and replaced by the new one (conversely deciding to keep the old JPEG object may imply having to archive the old version of the corresponding old MS Visio object as well).

Plan. The condition(s) and impact(s) of a change operation are connected to the `Dependency` concept in LRM via `precondition` and `impact` properties as illustrated in Fig. 4. These connect a `Dependency` to a `Plan`, *which is defined as a specialized description representing a set of actions or steps to be executed by someone/something* (*either human or software*); this is, thus, a means to give operational semantics to dependencies. Plans can describe how preconditions and impacts are checked and implemented (this could be for example defined via a formal rule-based language, such as SWRL). The temporally coordinated execution of plans can be modelled via *activities*. A corresponding `Activity` class is defined in LRM, which has a temporal extension (i.e. has a start and/or end time, or a duration). Finally, a resource that performs an activity, i.e. is the "bearer" of change in the ecosystem, either human or man-made (e.g. software), is represented by a class called `Agent`.

4.3 Domain Ontologies

A *domain ontology* (or *domain-specific ontology*) is a formal description of modelling concepts in a specific domain in a structured manner. Three media domain ontologies have been developed within PERICLES, aimed at modelling digital preservation risks

for the three respective subdomains (DVA, SBA and BDA), via LRM-based constructs (c.f. section 3.1). The key notions adopted and extended by the LRM are:

- **Activity** - represents activities that may be executed during a digital item's lifespan. The media domain ontologies extend the Activity class, in order to model activities that are considered to be important for digital preservation processes (creation, acquisition, storage, access, display, copy, maintenance, loan, destruction, etc.).
- **Agent** - with subclasses for human and software agents. Human agents are in addition specialised for the media domain into artists, creators, programmers, museum staff, etc. and software agents into programs, software libraries, operating systems, etc.
- **Dependency** (c.f. Section 4.2) - indicates the association or interaction of two or more resources in the domain ontology. For the media domain ontologies, we extend the basic notion of LRM dependency with three sub-categories:
    - **Hardware Dependencies** - specify hardware requirements for a Resource in order for it to function properly.
    - **Software Dependencies** - indicate the dependency of a Resource or Activity on a specific software (Software Agent) - name, version, etc.
    - **Data Dependencies** - imply the requirement of some knowledge, or data or information, in order for a Resource to achieve its purpose of existence or function. This kind of data may originate from human input (e.g. passwords), computer files (e.g. configuration files), network connection, live video, etc.

  The context of dependencies may be additionally enriched with the notions of intention and specification (see Section 4.2). For the media domain ontologies, a set of predefined intention types were defined:
    - **Dependencies with a Conceptual Intention** are aimed at modelling the intended "meaning" of a resource (i.e. artwork) by its creator; according to the way he/she meant it to be interpreted/understood. For example, a poem (digital item) belonging to an archival record may not conserve its formatting during the normalization process, something that may be contrary to the intention of the poet regarding the way that the poem is conceptualised/conceived by a reader.
    - **Dependencies with a Functional Intention** represent relations relevant to the proper, consistent and complete functioning of the resource. For example, a specific codec is required to display a digital video artwork.
    - **Dependencies with a Compatibility Intention** model compatible software or hardware components which may operate together or as replacement components for availability, obsolescence or other reasons. For example, the software used for playing back a digital video artwork consistently is compatible with certain operating systems.

A domain-specific instantiation, presented in Fig. 5, describes the following scenario taken from the BDA subdomain: a normalisation activity is applied in a text file (item, digital resource), according to the archival policy defined by the used normalisation software (OpenOffice). In terms of the media ontology, there is (a) a data de-

pendency of the normalisation activity to the item itself, (b) a software dependency of the normalization activity to the used software, and (c) a hardware dependency of the normalization software to the hardware required in order for the software to run efficiently. The intention of all three types of dependencies is functional, meaning that all the required resources modelled in this example impact the functionality of the resources for which the dependencies were implemented.

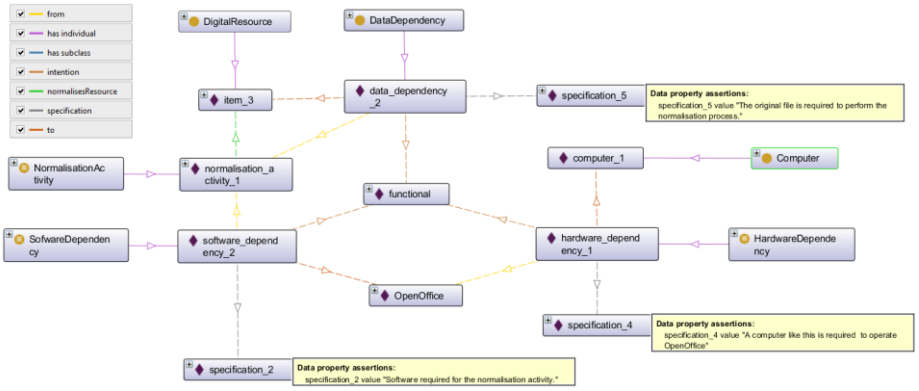

Fig. 5. Dependencies existing within the context of a normalisation activity applied in a digital item.

Within the context of PERICLES and the DVA ontology, an Ontology Design Pattern (ODP) for representing digital video resources was introduced [6]. This work was motivated by the problem of consistent presentation of digital video files in the context of digital preservation. The aim of this pattern is to model digital video files, their components and other associated entities, such as codecs and containers (Fig. 6). The proposed design pattern facilitates the creation of relevant domain ontologies that will be deployed in the fields of media archiving and digital preservation of videos and video artworks.

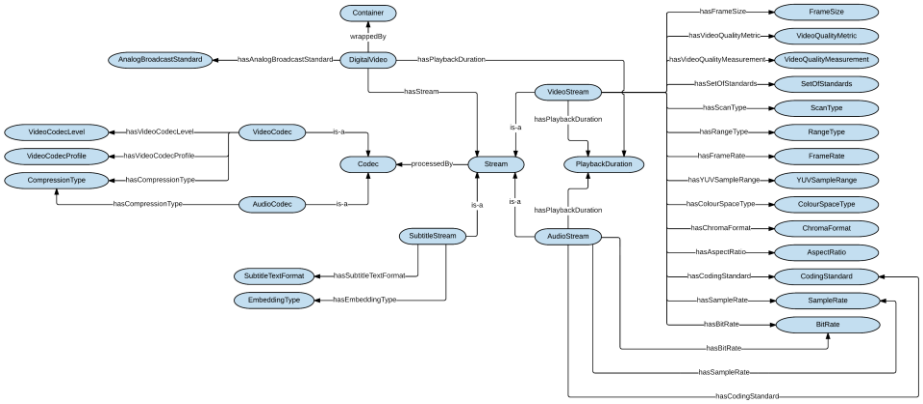

Fig. 6. Digital Video ODP schematic view.

The design pattern illustrates a more general principle, namely that ecosystem models can be constructed from a set of common templates. Such an approach would greatly reduce the effort required to create models by enabling a more modular approach where templates are reused across many different models.

4.4 Environment Capture and Model Population

An important consideration for a model-driven approach to preservation is to minimise the effort required to construct the ecosystem models. The *PERICLES Extraction Tool (PET)* provides one approach. PET is an open source framework for the extraction of the Significant Environment Information of a digital object. Here significance is a positive number expressing the importance of a piece of environment information for a given purpose. The tool can be used in a sheer curation [23] scenario, where it runs in the system background and reacts to events related to the creation and alteration of digital objects and the information accessed by processes, to extract environment information with regard to these events. All changes and successive extractions are stored locally on the curated machine for further analysis. A further extraction mode is the capturing of an environment information snapshot, which is intended for the extraction of information that does not change frequently.

The tool aims to be generic as it is not created with a single user community or use case in mind, but can be specialised with domain specific modules and configuration. PET provides several methods for the extraction of SEI, implemented as extraction modules as displayed in Fig. 7. The configuration has to be done once, after that PET can run automatically in a way that does not interfere with system activities and follows the sheer curation principles. The PET tool can be used to generate models based on the LRM ontology, which can then be used for ecosystem analysis. PET has been released as open source software under the Apache license. PET source code together with documentation and tutorials are available for download on Github[18].

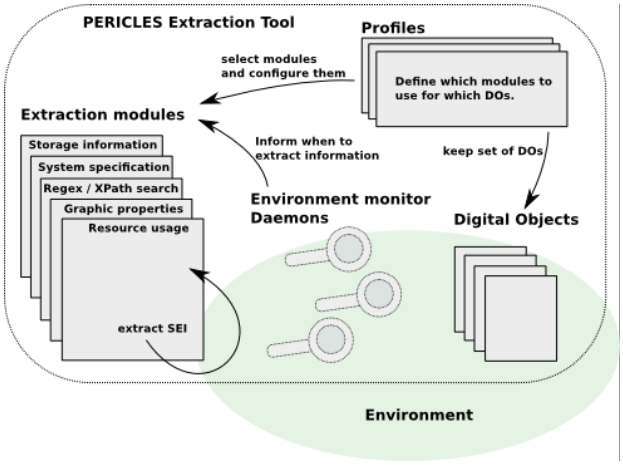

Fig.7. SEI extraction with the PERICLES Extraction Tool.

[18] https://github.com/pericles-project/pet

5 Applications in Digital Preservation

5.1 Monitoring of User Communities

One complex issue facing PERICLES is tracking the use of media (images or video) by different user communities. Although it is possible to apply usage tracking to a web portal making archival content available, once content is downloaded its use can no longer be monitored. Even this presupposes that portal users are registered and provide details of their intended usage, which may also not be feasible. What would be helpful in this context is the possibility to embed metadata and identifiers that would allow mapping and monitoring the diffusion of the media across user communities, in order to help identify their evolution.

Information encapsulation (IE) methods can be distinguished into the categories of packaging and information embedding. Packaging refers to the aggregation of files or other information formats as equal entities stored in an information container. In contrast to this, information embedding needs a carrier information entity (file/stream) in which payload information will be embedded.

The *PERICLES Content Aggregation Tool (PeriCAT)* [28] is a framework for Information Encapsulation techniques. It integrates a set of information encapsulation techniques from various domains, which can be used from within the framework. Furthermore PeriCAT provides a mechanism to capture the scenario of the user, and to suggest the best fitting information encapsulation technique for a given scenario. The tool is available to download on Github[19] under an Apache Version 2.0 licence.

5.2 Quality Assurance

Quality Assurance is defined by Webster[20] as "*a program for the systematic monitoring and evaluation of the various aspects of a project, service, or facility to ensure that standards of quality are being met*".

In PERICLES we define a series of Quality Assurance (QA) criteria for the entities of evolving ecosystems, in particular for policies, processes, complex digital media objects, semantics and user communities. This will allow us to manage change in the ecosystem by validating its entities, detecting conflicts and keeping track of its evolution through time. Our methods focus on validating the correct application of policies to the ecosystem. When change occurs, the approach will ensure that policies are still correctly implemented by tracing the correct application of the higher-level policies (guidelines, principles, constraints) in the concrete ecosystem implementation. Policies will be expressed at different levels, using a policy model integrated in the Ecosystem Model itself. We support the QA of policies by defining criteria and methods that can validate or measure the correct application of policies through processes, services and other ecosystem entities, so ensuring that the implementation is respecting the principles defined in the high-level policies. The QA methods in turn support the management of change in the ecosystem entities, such as change in policy, policy

[19] https://github.com/pericles-project/PeriCAT

[20] http://www.merriam-webster.com/dictionary/quality%20assurance

lifecycle, change in the processes implementing those policies, or change in other policy dependencies. In this model, we do **not make any strong assumption about the format in which the policy has to be expressed**, be it natural or formal language, nor are we imposing any specific structure on the processes used to implement them (although we are providing exemplar implementations). We are aware that policies and processes in real systems will be implemented using a variety of techniques and we aim to develop a policy layer that can be applied on top of existing ecosystems. This assumption will allow the deployment of such QA methods in systems that are not built using only specific technologies or rule languages, making their adoption simpler.

Other projects have looked at the issue of QA in preservation, in particular the SCAPE project[21], with a focus on the QA of a specific type of digital object (image, audio, e-publications), and also on the implementation of digital preservation policies by collecting metrics on collections; this is a valuable approach that is specific to issues related to digital object QA. Our approach works at the model level and addresses the implementation of digital preservation policies in existing ecosystems, although it takes into account the valuable work done by SCAPE.

More concretely, in [8] we define a policy model, a policy derivation process with guidelines, and a series of QA criteria and change management approaches for policies, taking into account the possibility of conflicting policies. We are currently working on exemplar implementation and refinement of the methods, implemented using the LRM and digital ecosystem models and other PERICLES technologies.

5.3 Appraisal

Appraisal is a process that in broad terms aims to determine which data should be held by an organisation. This can include both decisions about accepting data for a collection or archive (e.g. acquisition) as well as determining whether existing data in an archive or a collection should be retained.

In traditional paper-based archival practice, appraisal is a largely manual process, which is performed by a skilled archivist or curator. Although archivists are guided by organisational appraisal policies, such policies are mostly high-level and do not in themselves provide sufficiently detailed and rigorous criteria that can directly be translated into a machine executable form. Thus, much of the detailed decision-making rests with the knowledge and experience of the archivist or curator.

With the increasing volumes of digital content in comparison to analogue, manual appraisal is becoming increasingly impractical. Thus there is a need for automation based on clearly defined appraisal criteria. At the same time, decisions about acquisition and retention are dependent on many complex factors. Hence our aim here is to identify opportunities for automation or semi-automation of specific criteria that can assist human appraisal.

In [8], we set out our overall approach to appraisal. In Appendix 1, we extend the categorisation of appraisal criteria in the DELOS project. The latter focused on appraisal as the determination of the worth of preserving information, that is, as a means

[21] http://wiki.opf-labs.org/display/SP/SCAPE+Policy+Framework

of answering the question, 'what is worth keeping'? Here appraisal is considered as a process to be revisited throughout the life of the digital object. Consequently, our results differ principally in the breadth of material considered and in the number and breadth of appraisal factors identified.

Within the context of the PERICLES case studies, appraisal can naturally be partitioned into two distinct categories.

- **Technical Appraisal** – decisions based on the (on-going) feasibility of preserving the digital objects. This involves determining whether digital objects can be maintained in a reusable form and in particular takes into account obsolescence of software, formats and policies.
- **Content-based (or intellectual) Appraisal** – acquisition and retention decisions or assignment of value based on the content of the digital objects themselves.

Our focus in this paper is on the technical appraisal aspect. We are primarily interested here in predictive rather than reactive approaches to modelling the impact of change. Projects such as PLANETS [29] used a *technology watch* to detect changes in the external environment, which could then result in changes to archived content. We aim to model risks through understanding longer-term trends to predict the impact of changes in the future. This work follows a number of steps:

- Quantify primary risks to the ecosystem. This is done by analysis and modelling of external data sources to predict the likely obsolescence of software and formats, or hardware failure of entities in the ecosystem.
- Determine the impact of primary risks on entities in the ecosystem. This step aims to identify entities at the greatest primary risk.
- Determine the impact resulting from higher-order risks propagating through ecosystem. In this step we propagate risks through the models.
- Determine potential mitigating actions and their associated costs. In some cases, it may be possible to execute and validate mitigating actions automatically.

The overall goal is to provide a tool for use e.g. by archivists, to analyse a digital ecosystem, determine at what point in the future there is a significant risk to reuse, and the potential cost impact and potential mitigating actions. Such a tool could be applied for example to assess the value of a software-based artwork, by determining how long it can be displayed in exhibitions before elements become obsolete or require refactoring, or the cost of maintaining a set of scientific experiments for a given time period.

In order to enhance the user experience of the model-driven approach, PERICLES is developing a visualisation tool MICE (Model Impact and Change Explorer), which aims to present risk and impact information to users.

6 PERICLES Integration Framework

The PERICLES integration framework, described in detail in [9], is designed for the flexible execution of varied and varying processing and control components in typical preservation workflows, while itself being controllable by abstract models of the overall preservation system. It is the project's focal point for connecting tools, models and

application use cases together to demonstrate the potential of model-driven digital preservation.

The integration framework can be deployed in slightly different ways to suit testing and development, or real-world deployment. Fig. 8 shows the configuration for real-world deployment.

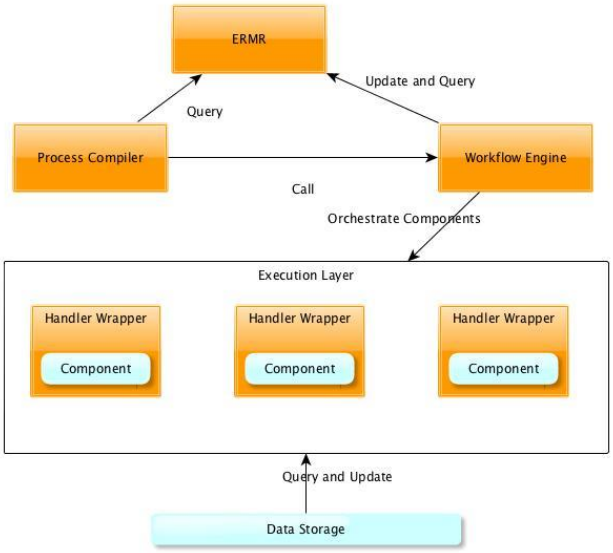

Fig. 8. Real world deployment of PERICLES components, based on the integration framework.

The *Entity Registry Model Repository (ERMR)* is a component for the management of digital entities and relationships between them. Access methods are presented as RESTful services. The ERMR registers and stores entity metadata and models. Agreed metadata conventions provide the necessary registry functionality; the registry is agnostic to the entities and metadata stored with the interpretation of data being the responsibility of the client applications. The ERMR provides a CDMI implementation for HTTP access to entities in the registry.

The ERMR uses a triple store as a database for the storage and retrieval of triples through semantic queries. The ERMR also provides a mediator service to integrate semantic services to extend its reasoning capabilities. It also provides a simple RESTful API for access to the registry. Queries can be expressed in the SPARQL query language to retrieve and manipulate data stored in the triple store. To link entities described as triples to actual digital objects, the ERMR is able to provide unique identifiers used to create a unique CDMI URL for an associated object. This URL is used to link entities stored in data storage components.

The *Process Compiler* (PC) supports the translation and reconfiguration of preservation process models described in the ERMR into executable workflows to be employed by the Workflow Engine. As part of this the PC will transfer information to the ERMR, which is used to update the process descriptions. The current component implementation is targeted towards a BPMN compilation system, though in theory the design can be adapted to any workflow engine language.

The *Workflow Engine* takes processes, compiled by the Process Compiler with descriptions and implementations stored in the ERMR, Data Storage and Workflow Engine cache, and executes them through orchestration of executable components (PERICLES tools, archive subsystems or supporting software) wrapped in Web Services Handlers with REST targets. This is one of the main fixed points of the architecture, which is active at all times.

Processing Elements (PEs) are any pieces of software or hardware that can be used by a PERICLES process to accomplish a given task. They are characterised by a fixed set of parameters including input and output types, platform requirements and versioning information.

Handlers are at the core of the integration framework. They function as the communication points for each major entity within the system. The Handlers deal with the validation of incoming requests, exercise the functionality of the PEs they wrap, store and transfer the results of PE functions and initiate necessary communications with other PEs. The Handlers do not perform any operations that change or alter the data contained in the objects they handle; only PEs can alter and change data. PEs, on the other hand, should not have knowledge of anything in their environment. This means they perform their function and only their function, and the Handlers deal with the rest of the PERICLES system and the outside world.

Data Storage is a specialised long-term Processing Element, a permanent service available to a PERICLES-based system. This component is responsible for storing digital objects, which can be data files, metadata, models or ontologies. Data Storage must be represented in the framework as a long-term service, since Processing Elements are typically transient in nature with limited functionality scope. The Data Storage service must manage data, as required, as bit-level preservation, object replication and distributed storage mechanisms.

7 Application of PERICLES Technologies to the Wider Community

As PERICLES is an inter-disciplinary project cutting across a number of technologies and aimed at diverse communities each with their specific remit, there are several approaches that the project is adopting for the exploitation of the results [1]. These include:

- Software products, e.g. component or system level modules.
- Services, e.g. on demand cloud-based preservation services.
- Consulting, e.g. advice on best practices and forming a preservation strategy.
- Training, e.g. commercial training.
- Education courses, e.g. Masters level courses taught at academic institutions.
- Technology licensing, e.g. through the use of patents.

The two application domains, space science and digital media are the primary targets for technology transfer due to the partner involvement and knowledge.

In space science, Europe (especially ESA) has made large investments over the years to develop, launch and operate missions in different fields of science (e.g. Earth

Observation, Planetary Science, Astronomy, ISS experiments). This has resulted in long time series and large volumes of data being requested and made available to different scientific users. However, no explicit mechanisms exist today to cover their maintenance long after the completion of the operations phase of the relevant missions and preservation is addressed diversely and only on a mission-per-mission basis. These data are unique and irreplaceable and constitutes a capital for Europe that is fundamental to generate economic and scientific advances.

The requirements for libraries, museums, galleries, and archives (and other heritage sectors) have evolved quite radically during the last fifteen years. Museums and galleries are increasingly collecting digital objects. These may be software-based artworks, design objects or digital objects related to the history of science and engineering. In the case of artworks, these will be acquired and preserved during their active life and will in the majority of cases evolve and change over time, for example for display in different operational settings. In other cases there may be a desire to keep a particular digital object functioning in a way that represents the historical context of that object. In either case, understanding and documenting what those objects are dependent on and how the digital environments and the objects change over time is essential to the mission of the museum. The user expectation is that this type of content will be permanently accessible and valuable. This type of demand is coupled with the expectation that archived content can be viewed in the "original form", independently of specific software or hardware technologies, thereby re-creating an "authentic" user experience, even if they are associated with software or hardware that is non-standard or obsolete.

Beyond the space science and cultural heritage sectors, media production is a growing sector. The business case for preservation in this environment is often based on the re-use of material in new productions; avoiding the expensive or at times impossible task of re-capturing material.

Digital library services provide the infrastructure to underpin teaching and learning; research and scholarly communication; web services; and other discovery services based on resource sharing across university and educational sectors. Solutions are required address the rapid growth and evolution of technology, formats, and dissemination mechanisms. Preservation requires the tools to provide access, support authenticity and integrity, and address the mitigating effects of technology or media obsolescence.

Projects in Science and Engineering are expensive to setup, or the situation for the project are unique. Funding agencies are increasingly seeking to ensure the data collected by these projects is kept long enough for any interested groups to make use of the data. For example, the UK Engineering and Physical Science Research Council (EPSRC) have started to require that collected data is kept usable for a minimum of 10 years[22], with many other funding agencies taking similar approaches. Although such policies exist, many of the funded research groups lack the expertise or the tools to meet this requirement.

Increasingly the area of healthcare is adopting ICT to store patient information and to aid in administering treatment[23,24]. By its very nature patient healthcare information

[22] http://www.epsrc.ac.uk/about/standards/researchdata/expectations/

[23] http://www.digitalpreservationeurope.eu/publications/briefs/security_aspects.pdf

is highly sensitive and subject to stringent policies that often differ across countries, or in some cases regions on how the data must be managed (including who can access the information and how it is disposed) making it very difficult, if not impossible, to distribute records electronically from one domain to another.

8 Conclusions

The current paper has provided an overview of some of the work being performed by the PERICLES project, at the end of the third year of the four-year project. The results we have obtained so far illustrate the potential value of models in digital preservation. The final year of the project will be focused on producing integrated prototypes and further developing the user-facing components such as appraisal.

The outcomes of the project align with the EU Digital Agenda for Europe[25] in supporting the digital preservation of digital cultural assets, and are potentially applicable across a wide range of sectors beyond the space science and cultural heritage.

References

1. PERICLES Consortium: Deliverable D10.1 – Initial version of exploitation Plan (2014). http://pericles-project.eu/uploads/files/PERICLES_WP10-D10_1-Exploitation_Plan-V1.pdf
2. Lagos, N., Waddington, S., Vion-Dury, J.-Y.: On The Preservation Of Evolving Digital Content - The Continuum Approach And Relevant Metadata Models, 9th Metadata And Semantics Research Conference (MTSR 2015), Manchester, UK. http://pericles-project.eu/uploads/files/PreservEvolvingDigitalContent-LagosWaddingtonVion-Dury-32-MTSR2015.pdf
3. Vion-Dury, J-Y, Lagos, N., Kontopoulos, E., Riga, M. Mitzias, P., Meditskos, G., Waddington, S., Laurenson, P., Kompatsiaris, I.: Designing for Inconsistency – The Dependency-based PERICLES Approach, First International Workshop on Semantic Web for Cultural Heritage (SW4CH 2015), Futuroscope, France. http://www.xrce.xerox.com/Research-Development/Publications/2015-052
4. PERICLES Consortium: Deliverable D2.3.2 - Data Surveys and Domain Ontologies, (2015). http://pericles-project.eu/uploads/files/PERICLES_WP2_D2_3_2_Data_Survey_Domain_Ontologies_V1_0.pdf.
5. Waddington, S., Tonkin, E., Palansuriyam, C., Muller, C., Pandey, P.: Integrating Digital Preservation into Experimental Workflows for Space Science, PV2015, Darmstadt, Germany. http://pericles-project.eu/uploads/files/PERICLES_PV2015_KCL_Presentation.pdf.
6. Mitzias, P., Riga, M., Waddington, S., Kontopoulos, E., Meditskos, G., Laurenson, P. and Kompatsiaris, I.: An Ontology Design Pattern for Digital Video, Proceedings of the 6th Workshop on Ontology and Semantic Web Patterns (WOP 2015) co-located with the 14th International Semantic Web Conference (ISWC 2015), Vol. 1461, Bethlehem, Pensylvania, USA.

[24]http://www.ombudsman.org.uk/__data/assets/pdf_file/0016/24631/Digital-Preservation-Policy.pdf

[25] https://ec.europa.eu/digital-agenda/en/digitisation-digital-preservation

7. Corubolo, F., Eggers, A., Hasan, A., Hedges, M., Waddington, S, Ludwig, J.: A pragmatic approach to significant environment information collection to support object reuse, iPres 2014, Melbourne, Australia. http://pericles-project.eu/uploads/files/ipres2014_PET.pdf
8. PERICLES Consortium: Deliverable D5.2 – Basic tools for Digital Ecosystem management (2015). http://pericles-project.eu/uploads/files/PERICLES_WP5_D5_2_Basic_Tools_for_Ecosystem_Management_V1_0.pdf.
9. PERICLES Consortium: Deliverable D6.4 – Final version of integration framework and API implementation (2015). http://pericles-project.eu/uploads/files/PERICLES_WP6_D64_Final_Version_of_framework_V1_0.pdf.
10. Higgins, S.: The DCC Curation Lifecycle Model. International Journal of Digital Curation. 3, (1), 134–40 (2008). Available from: http://ijdc.net/index.php/ijdc/article/view/69.
11. Ball, A.: Review of Data Management Lifecycle Models (version 1.0). REDm- MED Project Document redm1rep120110ab10. Bath, UK: University of Bath. (2012). http://opus.bath.ac.uk/28587/1/redm1rep120110ab10.pdf .
12. Committee on Earth Observation Satellites Working Group on Information systems and Services (WGISS): Data Life Cycle Models and Concepts CEOS 1.2 (2012). http://ceos.org/document_management/Working_Groups/WGISS/Interest_Groups/Data_Stewardship/White_Papers/WGISS_DSIG_Data-Lifecycle-Models-And-Concepts-v13-1_Apr2012.docx.
13. CCSDS - Consultative Committee for Space Data Systems: Reference Model for an Open Archival Information System (OAIS), Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2 (2012).
14. Upward, F.: Structuring the records continuum (Series of two parts) Part 1: post custodial principles and properties. Archives and Manuscripts. 24 (2) 268-285 (1996). http://www.infotech.monash.edu.au/research/groups/rcrg/publications/recordscontinuum-fupp1.html.
15. McKemmish, S.: Placing records continuum theory and practice. Archival Science Volume 1, Issue 4, 333-359 (2001).
16. PERICLES Consortium: Deliverable D5.1.1 – Initial Report on Digital Ecosystem Management (2014). http://pericles-project.eu/uploads/files/PERICLES_D5_1_1-Preservation_Ecosystem_Management_V1_0.pdf
17. Tzitzikas, Y.: Dependency Management for the Preservation of Digital Information. Database and Expert Systems Applications, pp. 582-92. Springer Berlin Heidelberg (2007).
18. Tzitzikas, Y., Marketakis, Y., Antoniou, G.: Task-Based Dependency Management for the Preservation of Digital Objects Using Rules. Artificial Intelligence: Theories, Models and Applications, pp. 265–74. Springer Berlin Heidelberg (2010).
19. Marketakis, Y., Tzitzikas, Y.: Dependency Management for Digital Preservation Using Semantic Web Technologies. Int. Journal on Digital Libraries 10(4): 159-77, (2009).
20. Vorontsov K. V., Potapenko A. A.: Additive Regularization of Topic Models. Machine Learning. Special Issue “Data Analysis and Intelligent Optimization with Applications”, 101(1), 303-323, (2015).
21. Wittek, P., Darányi, S., Liu, Y.H.: A vector field approach to lexical semantics. In Proceedings of Quantum Interaction-14, Filzbach (2014).
22. Wittek, P., Darányi, S., Kontopoulis, S., Moysiadis, T., Kompatsiaris, I.: Monitoring Term Drift Based on Semantic Consistency in an Evolving Vector Field. In Proceedings of IJCNN-15 (2015). http://arxiv.org/abs/1502.01753.
23. Hedges, M. and Blanke, T.: Digital Libraries for Experimental Data: Capturing Process through Sheer Curation. In Research and Advanced Technology for Digital Libraries (pp. 108-119). Springer Berlin Heidelberg (2013).
24. Geoffrey I. Webb, G.I., Hyde, R., Cao, H., Nguyen, H-L., Petitjean, F.: Characterizing Concept Drift. Accepted for publication in Data Mining and Knowledge Discovery on December 10, 2015. At http://arxiv.org/pdf/1511.03816v5.pdf

25. Darányi, S., Wittek, P., Konstantinidis, K., Papadopoulos, S.: A potential surface underlying meaning? Yandex School of Data Analysis Conference, Machine Learning: Prospects and Applications, Berlin (2015). At https://www.youtube.com/watch?v=PEnRqv9hyzg.
26. Lerman, K., Galstyan, A., Ver Steeg, G., Hogg, T.: Social Mechanics: An Empirically Grounded Science of Social Media. In Proceedings of International AAAI Conference on Web and Social Media Fifth International AAAI Conference on Weblogs and Social Media (2011). http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/3836/4393.
27. PERICLES Consortium: Deliverable D2.3 – Media and Science Case Study Functional Requirements and User Descriptions (2014). http://www.pericles-project.eu/uploads/files/PERICLES_D231_Case_studies-Functional_Requirements_gathering.pdf.
28. PERICLES Consortium: Deliverable D4.2 Encapsulation of Environmental Information (2015). http://www.pericles-project.eu/uploads/PERICLES_WP4_D4_2_Encapsulation_of_Environmental_Information_V1.pdf.
29. Aitken, B., Helwig, P., Jackson, A., Lindley, A., Nicchiarelli, E., Ross, S.: The Planets Testbed: Science for Digital Preservation, The Code4Lib Journal, Issue 3, (2008). http://www.dcc.ac.uk/resources/briefing-papers/technology-watch-papers/planets-testbed#sthash.mRsLQkcg.dpu.f

Datapipe: A Configurable Oil & Gas Automated Data Processor

Florent Bourgeois[1,2] and Pierre Arlaud[1]

[1] Actimage GmbH, Hafenstraße 3, 77694 Kehl, Germany
[2] Université de Haute Alsace, MIPS EA 2332,
12 rue des Frères Lumière, F-68093 Mulhouse, France
{florent.bourgeois, pierre.arlaud}@actimage.com,
florent.bourgeois@uha.fr

Abstract. Exploration and Production companies need to know where are the oil and gas reservoirs, how much they hold, and whether they can profitably produce oil and gas. Data collection, management and analysis are therefore central to the industry. As in most application areas, raw data are processed, implying several tools and experts interactions. Nevertheless, the oil an gas sector data processes imply unusual scale of, multimodal and long-lived, data alongside with complex analysis. DataPipe is a research project funded by the Eurostars Program of the European Commission which purpose is to develop a platform, toolkit and pipeline for the intelligent, rule-based selection, management, analysis, publishing and display of heterogeneous multimodal data in the oil and gas sector. This paper describes Actinote 4.0, a flexible web-based platform, which is developed to respond to the specified Datapipe context and is dedicated to the creation of specific domain-based process assistant applications that are certified by expert systems.

1 Introduction

Oil and Gas (O&G) are currently worldwide primary resources. Exploration companies perform seismic surveys to locate reservoirs and interpret physical properties of the rock. These surveys generate data which have then to be analysed in order to define the profitability of the reservoir. They can be complemented by aerial or satellite photography, gravitational measures. Plus, test borehole can be drilled to bring data with with positional, radioactivity, temperature, porosity, resistivity and other measures that enhance the geological model. Surveys have been made for years, implying technology evolution. The retrieved data have then changed and their storage also did. Improvements in analytic and drilling techniques as well as shifts in the global economy [1–4] can change decisions, so survey data have a long shelf life. A deposit that was once uneconomic may require a new analysis and interpretation thirty years later. There are problems of maintaining, tracking and accessing very large data volumes for decades, finding and mining old data from different storages, reading and interpreting different media formats and file types. The data are then multimodal, long-lived and difficult to manage. The expected valuable lifetime of seismic data ranges from four to twenty years - up to the point where newer seismic technologies make it cheaper to re-survey

than reprocess the data. As a result, data from many different types, formats and locations have to be found, managed and processed.

Over the past 20 years, geologists have built well databases, geophysicists have developed ways to handle seismic data and reservoir engineers have managed production data. There has been no common approach and the separate initiatives have created solutions with dedicated tools and procedures. Nevertheless, the current trend is to is to move from unstructured collections of physical and digital data toward structured sets of digital exploration, drilling and production data.

Big data and High Performance Analysis are now terms commonly employed to refer to O&G data management in industry white papers [5, 6], and scientific communities[7]. This is due to the massive amount of data it represents, the complexity of the algorithms employed and the variety of data format.

Within this context, Actimage, Dalim Software GmbH, Ovation Data Services Inc. and Root6 Ltd initiated a collaboration around the DataPipe project, which has been selected and funded under the European program EUROSTARS. DataPipe aims for proposing a solution to easily create automated data processor in the domain of O&G which can answer these different prerequisites. The project goal is to develop a platform, toolkit and pipeline for the intelligent, rule-based selection, management, analysis, publishing and display of heterogeneous multimodal data in the oil and gas sector. It will create a flexible system to provide web-based visualisation and decision support based on the analysis of extremely large datasets. The platform will be extensible to big data mining, analysis and display in a wide range of industrial sectors.

1.1 Project Goals

DataPipe will create a new approach to multimodal data management, data mining and presentation, based on process modelling and metadata-based process automation. A new methodology has to be implemented, in order to manage the enormous data volumes, the range of asset types and the processes applied to them in the oil and gas exploration sector. Based on research advances in a number of domains, the DataPipe platform comprises the following elements:

- Intelligent data workflow tools to control the selection and flow of data from multiple sources, their processing and publication on multiple platforms.
- A data selection and management framework, which will deal with the connection to multiple data stores across different APIs and ETL systems.
- A DataAgent processing platform, to connect to archiving systems and tape robots, with a Hierarchical Storage Management (HSM) system to bring files back from storage just in time for processing.
- Access control and security systems to protect the integrity of data from unauthorized personnel or attack to the data store.
- A cross-platform information display system to receive information from the framework and instructions from the job ticket, to tailor the publication.

To implement all of these goals is unsustainable for any of the project partners. Indeed, they require expertise and research specific to too many different domains. This is why each of the firms of the project has to handle a specific part according to their affinities with its themas.

1.2 Collaborating Partners

Being a firm proposing safe and reliable storage systems, *Ovation Data Services Inc.* has specialised in data management for the full life cycle of exploration data in the exploration and production industry. Their services cover all forms of seismic and well log data transcription and migration, plus format conversion, recovery, remediation and de-multiplexing. Their role in the project is to define the domain specific expertise, the existing processes and their results and provide expert feedbacks. Aside of this, they will have to furnish web services to access to their services in order to handle the data all their stages (acquisition, processing, storage, archive, deep storage, destruction) and specify their access requirements.

Root6 Ltd ContentAgent systems address workflow problems in the (digital) film, video and advertising industries. The are designed to support multi-format video media encoding, faster-than-real-time transcoding and streaming transcoding. Their expertise in multiple format video transcoding processes automated and parallelized through the creation of complex configurable workflows is currently limited to local processing. Through this project, they aim to extend the system for Cloud-based operations and adapt their system on the multiple format data found in the domain of O&G to provide faster and automatized data processing based on parallel computing agents.

Dalim Software GmbH specialises in software systems implementing JDF for print, in which XML elements describe all the production processes and material types, regardless of the individual workflow. Their solution enables the creation of processes combining user steps (approval), technical steps and stage steps (milestones) with relevant results stored on databases accessible anywhere through web interface based on the user identified role. They bring to the project their expertise in processes designs implying automated steps and user interactions enabling complex processes requiring human supervision and decisions.

Actimage has been involved in many complex online projects involving cloud-based architecture and has wide experience of combining different mobile technologies. The existing Actinote system is designed to gather and deliver user- and context-centric information for mobile professionals. The system employs scalable Cloud architecture, interacting with multiple data sources (ERP, PIM, etc.), a blackboard multi-agent system to handle data and provide security, delivery and publishing of user related content. Our systems provide a solid expertise for the delivery, security, multi-platform and intelligent interface aspects of DataPipe. Based on this, plus our expertise in mobile and recent web technology applications and requirement based solutions for professionals, our role in the project is to implement user interfaces for the creation of workflows coherent with the specific domain of application that the O&G context represent, mobile devices application for the execution, supervision of decision support based workflows with means of publishing and displaying the information in format tailored to the user needs.

This paper focuses on Actinote 4.0, the solution Actimage developed in the defined context. First it details the goals and requirements specific to the development of a platform for the creation of domain specific based workflow applications. Afterwards it presents the Model Driven Engineering (MDE) approach the platform is founded on. There will then be a presentation of the implementation choices to implement the

method into the Actinote 4.0 platform and a demonstration of the extensions required to implement a reduced O&G data management as a domain specific case to evaluate the solution. Eventually the paper concludes on a review of the results of the project and further work perspectives.

2 The Solution Specification

The previous section highlighted the complexity of O&G data management and the need for a solution to standard their processing which includes complex and time consuming computer assisted activities alongside few user decisions making. This section details the requirement the solution must fill through a stakeholder point of view and then explains how such a specific domain solution can be abstracted to a multi-domain data processing solution addressing a wider range of applications.

2.1 Requirement Analysis

In order to propose such a solution it is important to point out the different actors that are likely to interact together for the generated applications lifetime. We identified four main users in this context.

Software Development Entity. Actimage, as a software development entity aims to propose a solution responding to the domain needs. This implies the generation of application able to handle O&G data by retrieving them from their multiple identified storages and to process them according to data management expert in order to then apply O&G experts analyst specific processes. These processes then result into bankable data that have to be reported and stored for later use.

Considering the number of experts, processes, data storages and data formats, it is not possible to create one application able to respond to any of the processes. On the other hand, it is unthinkable to produce specific development projects for every processes. It would lead to the production of an unlimited number of projects in order to propose to each user the specific process they want to execute with their specific requirement concerning data storages and expected results. It is then obvious that Actimage requires the development of a solution implying an application editor based on software engineering to enable fast and easy development of data processing applications. Thus Actimage will be able to software engineering solutions based on the specific activities of users instead of constraining them to a rigid generic process.

End-users. The end-user brings to the application creation process its expertise on his specific domain of activity, his habits and his expectations on the look and feel of the application. The application process must assist the user in his task and then be a support in his common activities and not a burden. It is then important to give the software architect tools to answer the user needs. It is important to note that the end-users expertise implies specific business vocabulary. hence a gap between the software architect which manipulates software artefacts and the end-user languages naturally

arises. The solution then have to reduce this semantic gap in that it proposes to the software architect artefacts corresponding to the end user's vocabulary. The end-user describes a data management process. This description contains the specific data he wants to use, which implies their format and location, the different algorithms to apply, the decisions to take according to the data and the expected output and output format. A data processing assistant like this becomes pertinent on smartphones if the processing can be outsourced on a distant powerful network of computer and if the application is able to display the pertinent results, enables user decisions and assignation of tasks to other users. Plus if the application present ways to assign tasks to other users, it would be possible to create processes with several end-users interacting to combine different expertises on the data. This requires the solution to handle device to device communication for the assigning and the data sharing activities.

Data Management and O&G Specialists. It is important to note that the generated process if not correctly described can lead to not applicable process in the domain. Indeed, the end-user is an expert in the analysis of specific data. But we mentioned in the previous section that before the data are in the analysis format, they have to be located and retrieved from several data storages, homogenized and made accessible. Due to the disparity of current solutions, O&G experts are used to handle such processes but it is a burdensome task without any of their expertise added value. On top of this, the algorithms of O&G experts analysis processes also imply a set of rules and conditions that only specialists knows. Besides, specification of a process can also entail errors.

Therefore, the different activities composing a full data management process imply activities which present constraints such as specific order (archive data must be the last activity), specific data type as input (only unarchived data are convertible).These specific domains activities being at the center of the processes, it is mandatory that the solution proposes mechanisms to handle them correctly. The specific domain knowledge must be embedded in the application creation process in order to assert that the generated processes are possible and manageable considering the limitation of the system and the sciences of the data management and analysis. Consequently, an analysis of the domain semantic must be performed on the process.

Data management and O&G specialists are then required to define the domain semantic and the set of rules to apply on the system. Their role is then the description of a prescriptive framework to strictly observe. Since several types of analysis can be processed by the different applications, depending on their expertise target, it is of first interest to propose the creation of rule libraries. Each library embeds a specific expertise domain tool, to enable the specific elements of the expertise in the required process creation and avoid to present the unnecessary elements.

Architect. At this point, Actimage solution enables the creation of correct applications in the domain of O&G data management and analysis through the use of an application editor. The person in charge of the application creation is called the architect. The architect has to follow the end-user application description in order to generate an application able to assist him and automize his activities.

The architect goes through two activities. First he engineers his editor environment choosing over the set of specialists libraries and importing the necessary ones. This enables the use of specific domains tools according to the end-user domain and needs. He then performs a process engineering activity creating the application within the editor according to the description given by the end-user. The solution has then to propose tools to compose processes implying end user centric and fully automatic steps. Both must present user interfaces, respectively one to enable user interaction and one for supervision, enabling the end-user to oversee automatic steps evolutions and anticipate upcoming steps.

The different actors interactions described are illustrated in the figure 1.

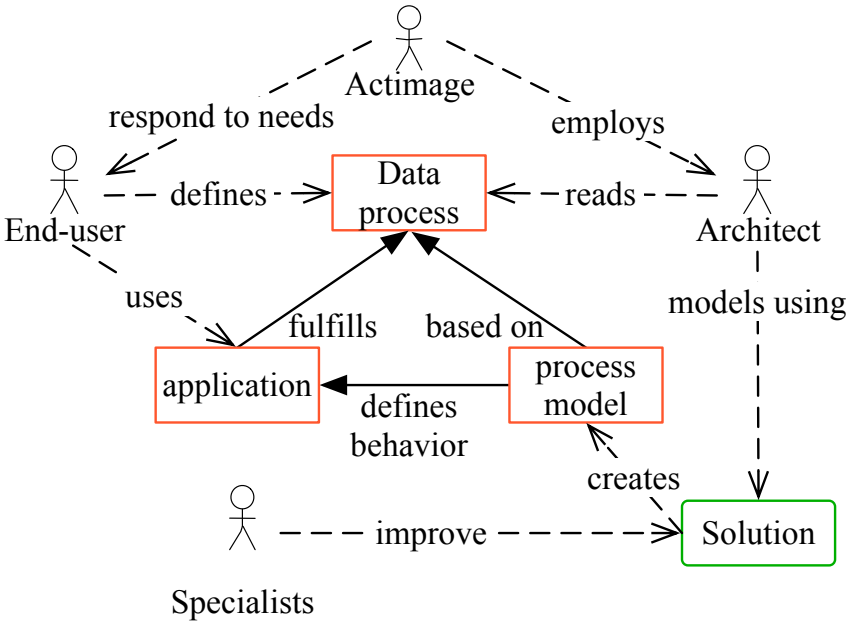

Fig. 1. Solution actors interactions.

To resume, the solution has then to address the following requirements:

- Definition of a complex process composed of automated and user centric activities.
- Execution of process on mobile platforms.
- O&G based processes creation assistant and validation.
- Manipulation of heterogeneous data and data sources.
- Distant processes execution for optimisation and balance of device processing load.
- Role based task assignations to users.
- Web based user interface for the creation, assignation and supervision of processes.

2.2 Solution Generalization

Until now we defined that ACTIMAGE have to furnish tools to create O&G data management processes and to supervise the execution and results of such a process.

It appears from the requirement analysis that the processes are mainly focused on data management and manipulation. Providing that the communicating services are well handled and that the processes are validated by the data management and O&G semantic when created, the format of the data in the application is coherent and the user can manage his work without worrying about the data coherence.

This solution is then based on two specific domains in order to automatize and simplify the expert knowledge based process creation. Since processes are implemented in supervised applications, it is easy to assert that the process status (created, running, terminated) and it asserts that, if the application is correctly described by the end-user, there is no step in the process that can be forgotten.

Such process based applications make sense not only in DataPipe's project context but also in various other domains. In custom-made industry, the prospect phase of meeting the client, understanding its wishes, capturing the environment and estimating the cost of the product would widely benefit of applications that assert that the whole process is performed, that simplify the collection of data either from the firm store to present the products or from the on site visit in order to capture the context of the sale. This domain presents the same requirement as O&G. There are heterogeneous data manipulation, such as camera pictures, measures, notes. Sensors, user interface, cloud-based database are the domain's multiple data sources. That's why all kinds of skills are employed, such as knowledge on the products sold, on the sensors use, or on the price calculation.

It is even imaginable to create applications meant to assist everyday life activities. Indeed, cooking, sport, handiwork are activities that can be represented as processes, requiring user interfaces to guide and assist them, which consume not only data but also materials and create results (meals, health status and manufactured furnitures respectively).

It is then possible to define a common metamodel the three cited examples correspond to which. This metamodel is illustrated on the figure 2.

Thus, instead of a solution based on the verification of two domains, Actimage proposes to define a solution for the creation of multi-domain heterogeneous data handling processes. The processes are validated by their coherence with the domains semantic.

The creation of the solution hence implies several modules:

- domain specific process creation editor.
- heterogeneous data manipulation and presentation interfaces.
- knowledge semantic modeling and verification.
- mobile device applications creation and execution.

The following section presents a MDE approach which was developped to design Actinote 4.0, the generic solution implemented to respond to the mentioned requirements.

3 Model Driven Engineering Approach

As explained in the previous section, the challenges of the Datapipe project can not be resolved by implementing applications for specific user cases unless a whole solution,

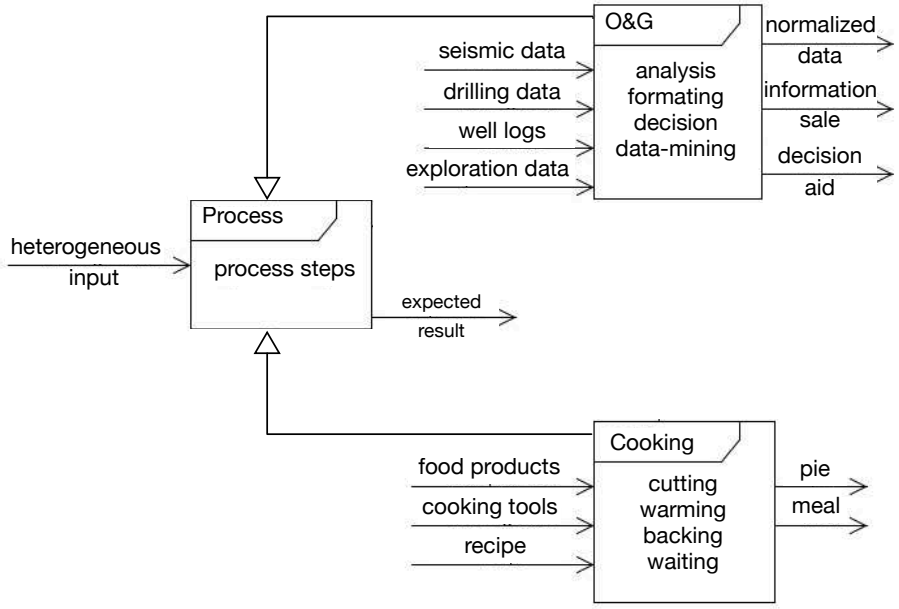

Fig. 2. Processes models and metamodel.

targeting the creation of domain specific data handling processes applications, is developed. This sections describes the structure and producing logic of Actinote 4.0, the solution developed by Actimage.

3.1 Background

Model Driven Engineering (MDE) is a software development method that considers models as the first class artefacts, even considering that everything is a model [8, 9]. Its purpose is to rely on models as development entities and then generate models of lower levels or even code, mapping between models abstractions, model evolutions, system behaviors or applications through the use of model transformations[10, 11].

MDE commonly defines models as a representation of an aspect of the world for a specific purpose. A model never represents the full system, but an abstraction of the system complete enough to represent all the required feature for a given use. A metamodel is a representation of a language able to describe lower abstraction level models. All the models described by the language are conform to the metamodel. This conformance relation thus asserts that the model is constrained by the semantic of its metamodel.

A model transformation takes a model as source and produces another model as target. A transformation metamodel is a mapping between the source model metamodel and the target metamodel.

Surveys [12, 13] proves that MDE, while being a more than ten year old method, is still a recognized method in software industry and several development teams use it in order to approach complex systems development.

OMG's Model Driven Architecture (MDA) [14] is one MDE initiative with a three layer structure. A Computational Independent Model (CIM) describes the business model (e.g. the UML grammar). Then it is transformed into a lower level model through the use of the language it represents. This generates a Platform Independent Model (PIM) which is in our example a specific model described with UML. Last, the PIM model is transformed into a Platform Specific Model (PSM). The generated PSM is the implementation of the system described by the PIM with technology specific to the targeted environment. In our example can be the android application code. Even though our approach does not matches exactly the MDA structure, we will use the CIM, PIM and PSM terms to identify the level of this paper upcoming models.

Using models to specify the system functionalities and then apply model transformations on them, so the implementation is generated, simplifies creation of a group of applications sharing the same description paradigm. It is possible to define a Domain Specific Language (DSL) which is a simple language optimized for a given class of problems[15]. This class of problems is named domain. A DSL enables an easily description of applications in a specific domain using a reduced set of elements. Since the language proposes a reduced set of elements, the model description and mappings are simplified compared to general programming languages, such as C++.

3.2 The Approach Global Structure

As described before, MDE approaches are based on models and their transformations to describe software behavior and automize their implementation based on this description.

Thanks to MDE it is possible to describe a DSL dedicated to the modeling of processes. This DSL is represented by a CIM model. Moreover, the architect editor is based on the DSL. This architect composes the application description with terms extracted from the DSL and thus creates the application process model. This model describes the functionality of the application without considering the implementation specificities. It is then a PIM model. A model transformation consumes the process model afterwards in order to produce the PSM corresponding application.

Nevertheless, the domain specific semantic brought by the specialists' knowledge is complex and implies deep modifications in the DSL with addition of domain specific terms for the architect and, more importantly, of semantical constraints that are hard to represent on models.

Indeed, constraints are often added to the modeling language by the addition of files containing the constraints' descriptions in text such as Object Constraint Language (OCL). This solution presents several downsides. Constraints are placed over objects and object relations, complex constraints are difficult to implement with this approach. This is a problem considering that the domains might be quite complex (e.g. relying on measure semantic). Also, they are in separated files that have to be updated in parallel with the model evolutions. Besides, the semantics have to be analysed on full processes that are only limited by the end-user's description.

In this fahsion, instead of dealing with cumbersome constraints programming, we chose to dedicate the semantic analysis to expert systems detailed later in this paper. We consider that the different domains semantics do not overlap.

To resume, our MDE approach's general structure is composed of three modules and two model transformations that are represented on figure 3. The editor enables the architect to model his processes with domain specific elements, the expert system analyses the domain specific semantics associated to the process to allow only the creation of coherent processes and eventually the application is the final distributed product with which the end-user interacts to execute his process.

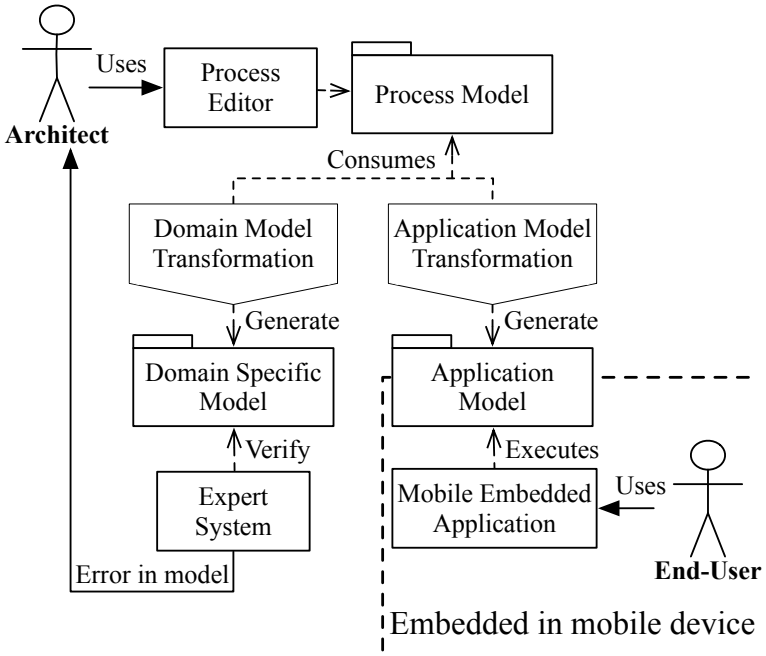

Fig. 3. MDE approach global structure.

This section now provides more details on the different modules composing the approach.

3.3 The Domain Specific Language Definition

In order to be able to create, represent and produce process-based editors and applications, we have to define a DSL able to describe all the possible processes models.

A standard description commonly used to model data processes is the workflow modeling paradigm. Workflows are defined as the automation of a business process presenting several activities, processing any kind of data and connected through transitions[16]. It is a widely used paradigm based on simple elements (activities connected through transitions) defined as being able to represent any kind of process[17]. A large community works on normative use of its elements[18, 19]. In our context, this abstraction can be used as metamodel used to produce the processes representing models.

Looking at figure 2, it is possible to make a direct parallel between the workflow activities, their inputs/outputs and our processes' metamodel. It is also possible to consider a user choice as an activity that transforms two potential futures processes' path

as the one that will be executed. The sole difference is that workflow activities are connected through transitions that can present conditions. These transitions conditions might be associated to either the presence of a correct data (resulting from an upstream activity) or user actions.

Using workflows as a standard representation for our processes' metamodel presents three major advantages. First, it is a simple abstraction that any software architect is used to manipulating, which makes the editor's main elements easy to assimilate. Second, it is possible to propose to the architect complex specific domain elements as simple activities or interfaces. This reduces the semantic gap between the end-user and the architect during the process modeling phase. This enables the creation of processes with less interactions with the end-user to require more precise description. Finally, a lot of workflows editors already exist, for example Datapipe partners already propose solutions based on workflows created through editors. Since those editors create models corresponding to the workflow metamodel, it is possible to use them as an editor approach. For this to be possible, the sole requirement is that the editor can be extended to provide the domain specific elements to the architect and also for the model transformations to be created. Figure 4 illustrates the impact of the use of several workflows editors on the transformation between the process editor and the expert system.

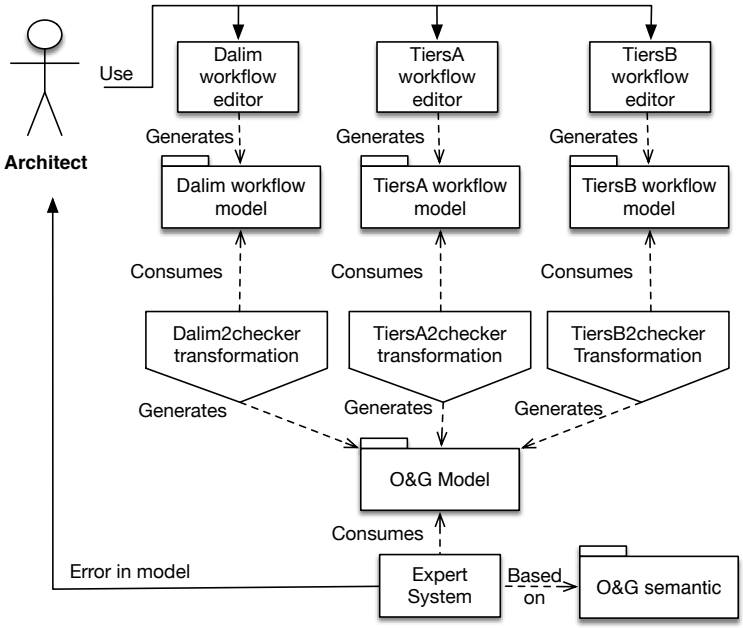

Fig. 4. Use of multiple editors.

3.4 Expert Systems

An expert system, or Knowledge-based System (KBS) is an AI System (input, transformations, output) with several blocks which understands expert knowledge and infers

behaviours to solve a problem in a specific task domain. Expert systems were already used in 1986[20]. They have matured over the years and are now still used especially with the rising domain of ontologies.
There are two types of knowledge[21]:

- Factual Knowledge: Deductions that an expert system should handle as is. Similar to the concept of axioms. This knowledge is widely shared and typically found in textbooks or journals.
- Heuristic Knowledge: The knowledge of good practice, good judgment, and plausible reasoning in the field. It is the knowledge that underlies the art of good guessing.

It is usually said that knowledge-based systems consist of two parts: a knowledge base and an engine. Therefore, as shown on the figure 5, the two basic generic blocks of an expert system respectively have these two responsibilities.

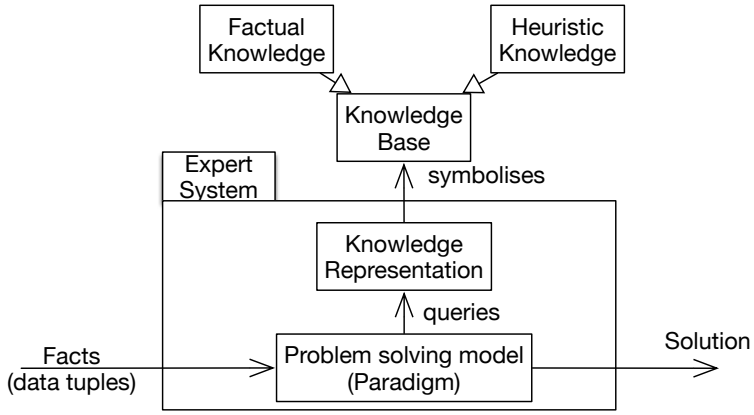

Fig. 5. An expert system.

Thanks to this approach, we are able to ensure that the semantic of the domains will be respected. Such a system, being able to check a domain, is also able to do so with several domains. We did not find examples of overlapping domains, hence our not considering issue of different domains interactions.

3.5 Application

Since our system is meant to generate a family of applications for dedicated domains, it is expected that a end-user requests several applications for different processes. Moreover, the process of posting applications on stores is cumbersome and there is no control over how to access these. The platform must provide a solution to ensure the privacy of the end-user's intellectual property to prevent any unauthorized access to the application, while ensuring the delivery of the application through a simple system.

With this MDE approach, instead of creating a new application for each process model, we propose to translate the model into a PSM which describes the expected behavior of the application and the different interfaces (graphical or to services) that it will use. Then, a unique application will handle any of the descriptions and, based on

an interpreter technology, will execute a behaviour corresponding to the descriptions. The descriptions can be sent to the applicatio through standard push methods on mobile connected devices, automatically providing the new process to the application once it has been created and verified. Combined to a login logic, this allows us to propose a unique application to group and execute any of the processes the end-user requests.

The interpreter executes the application's behaviour according to the workflow activity. Each activity is a milestone in the execution that either starts a process on data or request a choice from the user. The application then only requires to be able to read the workflow and compose interfaces according to the descriptions made in the models. That's why the application must know the editor's different elements in order to be able to interpret them on execution.

Hence, we can propose an application for the different existing mobile platforms which then can handle any of the process models. Which makes the approach able to target multi-platform mobile devices.

Through this section, we presented a Model Driven Engineering approach which enables the creation of a family of multi-domains well-founded processes applications. The next section presents an overview of Actinote 4.0, the Actimage implementation of such an approach.

4 Actinote 4.0 Implementation

The later section presented an MDE approach which answered the complex requirements of Datapipe project. This section presents the Actimage solution Actinote 4.0, implemented following the presented approach. Several details are considered to be out of the scope of this paper due to the industrial nature of the solution. This especially encompasses the different model transformations that will thus not be described.

4.1 The Editor

The editor is the module that is meant to be used by the architect to model the process executed by the application. We stated that such process implies graphical interfaces, workflow activities and data management. The lack of standard processes has encouraged our project team to give users a sense of intuitiveness in the way they can model their activities. The transfer of their operational process into the Actinote 4.0 platform is made accessible with a graphical approach: the architect extracts the flow of the end-user's process in a nodal diagram (which looks similar to a finite state graph) and the description of their constraints.

This normalization and meta-modelling ensures the reliability of the data stores. Not only will the homogeneity of this assemblage facilitate the computational discovery of patterns in the inputs, but it will also allow the utilisation of safeguards based on the specific domains. Since all inputs need to be specified and enumerated, there is in fact no way for the mechanism to be semantically ambiguous. Any incoherence can be spotted beforehand, insuring the integrity of the business knowledge.

The basic metamodel representing the editor DSL is illustrated in the figure 6.

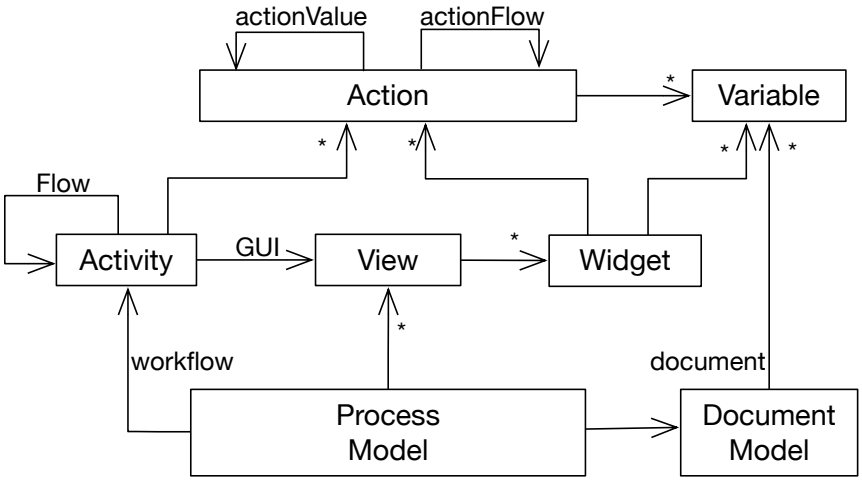

Fig. 6. Actinote 4.0 DSL's metamodel.

With the language represented by this metamodel it is possible to describe any processes. The workflow is modeled as activities chained together through flow links. Each activity presents a graphical user interface composed of several widgets. A widget is a graphical interface element that enables to give access (display and manipulation) to data or make decisions for the end-user. Variables are abstractions of the data manipulated by the process. The document model groups all the data that must be retrieved as the process result in order to automatically generate a report. An action is a specific domain process to apply on a set of data. Some common or complex actions, such as the retrieving of multi-storage data in O&G context, are added to the editor when a specific domain is imported. The architect also can implement specific actions with a nodal diagram dedicated for the data processing. Actions are started either by an activity or by user interactions on widgets.

During our test activities, we observed that the semantic gap reduction brought by the use of our DSL and the abstraction of complex processes as importable actions did not only help the architect understand the end-user specific domain vocabulary, but it also enabled the end-users to edit their own workflows. So our solution, as long as it provides the different complex operations of the process as element of the editor, enables the end-user to model his process model himself.

Therefore Actinote 4.0 is a good fit for the industry because it focuses primarily on the designing of forms and the web-visualisation of analysis results. The whole idea behind this work resides in the opportunity for an expert to be relieved of the time-consuming task load that converting data into a generic form can represent. Thanks to this effort, geologists, geophysicists and engineers can use the DataPipe platform and toolkit to publish and display heterogeneous multimodal data in their realm of expertise. The principle of this responsibility decoupling is that we separate the business logic of the process into three parts: the orchestration of its flow with the activities, the algorithmic aspect of each of its steps with actions and the designing of the display that will provide the users with a mobile access to the process with the widgets. It becomes

thus possible to partition the effort for different employees with different qualifications. Not only will domain-specific experts have the ability to engineer process for virtualizing and structuring production data without requiring any particular skills in software development, but the technical operators and decision makers will be able to run the predefined scenarios independently.

4.2 Workflow Validation

We mentioned in the previous section that the process models have to be validated in order to implement them into applications. This implementation approach adds several users and platforms-based validation requirements. Then, a workflow validation goes through multiples checks:

- Permission and access rights, which may require verifying the coherence of the rights.
- Semantic analysis of fields use and their types.
- Semantic analysis of domain specificities with expert system.
- Vacuity and halting tests (the workflow must have steps reaching a end).
- Responsiveness aberration tests for small displays.
- Consistency checks of the actions graph (which is a set of algorithmic nodes).
- Syntaxic analysis of the actions graph.
- Syntaxic analysis of the activities graph.
- Check of all unused elements (may they be variables types, resources, event graph parts).

Much of these requirements are resolved by the editor's language with typing of actions, variables and widgets and are out of the scope of this paper. We will only detail the expert system validation process.

Rules Engine Implementation of an Expert System. The Actinote 4.0 expert system encodes knowledge in first-order predicate logic and uses the Prolog language to reason about that knowledge. It hence uses a rules engine, which is the most common implementation of an Expert System and based on rules.

The knowledge is represented with a set of production rules Each data is matched to the patterns described by these rules with algorithms such as Rete Algorithm.

The solving entity is thus an inference engine (a.k.a. Production rules system), which uses either forward or backward chaining to infer conclusions on the data.

It's worth noting that, although conclusions are usually implied, their being here inferred shows that we indeed deal with Artificial Intelligence, so the system makes conclusions as humans would.

The figure 7 shows what the expert system thus becomes. Please notice that the Knowledge Base is not explicitly added on the diagram. The confusion between a knowledge base and the way it is represented in our system is usually made on purpose: in a rules engine, we call knowledge base the set of production rules, and not the actual knowledge that the experts have in their brains.

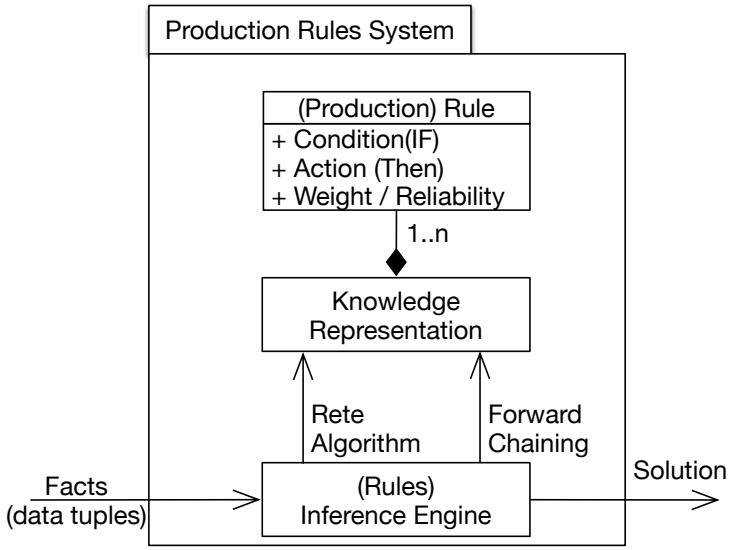

Fig. 7. An expert system based on forward-chaining.

4.3 Data Management

The first target for Actinote 4.0 is to answer to the aforementioned O&G domain data management complexity proposing a front-end to manage the enormous amount of data and their heterogenous nature.

Big-data Choice. Actinote 4.0 has a native support of MongoDB sets of databases. MongoDB is the most used documental database which puts the emphasis on multi-datacenter scalability, resulting in big-data model flexibility and performance. Big-data mining, analysis and display in a wide range of industrial sectors is made possible with this choice.

The data locality of MongoDB instances is an appropriate answer to the needs of O&G data management companies in terms of data. Not only does MongoDB handle billions of documents, but it also sustains hundred of thousands of database atomic operations per second, making it a suitable system for analyzing data. Since it?s also multiplatform, MongoDB can be scattered all across the globe to unite important seismic statistics and pieces of information.

Horizontal Scaling. All the unstructured collections of physical and digital data of the O&G data management companies may be dispatched in structured sets of exploration, drilling and production data. The data can then be split into different shards, meaning there will be different MongoDB servers for different ranges of data. For instance, one may divide the stores geographically and have non-overlapping immutable chunks for each predefined ?location? field corresponding to each area. Considering the built-in geospatial indexes in MongoDB querying system, exploring results of decades of tapes end other capturing data is ensured to remain performant.

Theoretically, there is always the possibility to include a Hadoop framework to solve storage and processing problematics in a distributed way. Computer clusters can thus be accessed to run complex analytics and data processing with Map-Reduce jobs.

4.4 Application

In accordance to the diversity of available media with the Actinote 4.0 software, it's also worth mentioning that it consists on mobile devices of a Qt application, which enables a good homogeneity of resulting behaviors on all platforms. The support of many features such as camera, contact list or network connections are handled in the same way on all platform and the compatibility on most devices (either on iOS and Android but also BlackBerry 10 for instance) remains assured. Another positive consequence of this choice is the integrated ergonomy of the OS: Qt framework adapts to the operating system it is running on so that it can use the standard approach for each graphical component. By doing so, the operators who are running the scenarios can keep the devices they are used to work on and we don?t have to handle resistance to change.

Network of Stores. The structured sets of data are organized in a web of servers and services which are all put together with the cloud computing procured by Actinote 4.0. The uniformity contract of the sets at our disposal can be made practicable by including converters and aggregators of data, or more generally ETL systems, all with the purpose that they are reunified in the beforementioned big-data schemes. In practice, one will firstly design a process, with the benefit webservices and ETL invocation. Secondly, the recorded knowledge will need to be digitized when no virtual save exists in computer understandable formats. Last but not least, this restructured aggregation will be merged and redistributed by means of sharding. This datamining process will maintain the sporadic existence of data with but two main differences: the data will be normalized so by construction rather easy to browse and the interface between all stores will be specified to ensure every piece of information is obtainable on the network. The ontological approach of metrology subjects is a good start for interpreting the production and exploration data which has been performed by Actimage[22].

4.5 Simple Oil & Gaz Implementation

We now present a simple example of the modules modifications implied by the use of a specific domain knowledge. Lets imagine that O&G processes are composed, instead of complex data management activities, of simply four different data manipulation operations: search, format, compare and store.

Search is based on a webservice which returns the data from a selected world area. It can return either numbers or string data depending on the world area (emulate different data storages).

Format allows to format a data into the requested format. If the selected data is already in the correct format, it simply does nothing.

Compress is able to compress a number data.

Store is used to create a uniformed storage. It requires that the data has been compressed.

We will call this domain small O&G in this example.

DSL Additions. In order to handle small O&G processes, the editor requires to propose to the user elements derived from our specification.

The data manipulation operations are actions. A specific action for each of the operations is added in the DSL. Since these actions are obviously distinct we also propose to create one type of activity for each operation. These activities will call the corresponding actions when the user validate their execution.

The Search activity proposes the architect to enter the name of the area in the world to search the data for and returns the result.

Format enables the architect to choose the variable to format, the expected output format (strings or numbers) and the variable in which to store it. It requires the output format and the variable types to be identical.

Compress let the architect choose the variable he wants to apply compression on. It requires the variable to be a number.

Store enables the architect to select the variable to store. It also allows him to choose a storage database to target. This parameter is shared between all stores.

The language also get two type of variables: string and numbers according to the manipulated types by small O&G operations. Compressed data is not a type of data because we consider since the compression is a non destructive operation.

The Expert System. Our expert system is fairly simple because the variable and activities typing validates most of the constraints brought by the domain semantic. Thought, the compressed status of data being not inferred in the process model it is the expert system role to handle it. Also, the type of the data returned by a search action has to he modeled in the knowledge base in order to assert that search actions stores data in corresponding type variable.

The knowledge base is then composed of facts concerning the search areas and rules to verify that search and compress activities are correct according to the expert knowledge. This knowledge being: search activity is always preceded by a compress activity and store activities variable type and storage returned data must correspond. Listing 1 shows the prolog knowledge base. The transformation of process model to specific domain model consumes a small O&G process model and generates a knowledge base extension which contains the different activities, actions and variables facts:

Application Modifications. In order to be able to execute small O&G processes, the application must be upgraded. It is mandatory to provide to the application the code to execute when actions are executed. The GUI widgets composing the small O&G activities are standards validation and data display widgets, hence there is no further development required to adapt the application to the new specific domain.

Listing 1. Small O&G knowledge base.

```
dataArea(string, asia).
dataArea(string,europe).
dataArea(number,America).
dataArea(string,Affrica).

searchVerification(ID) :- searchActivity(ID, Area, VarID),
           variable(VarID,Type),
           dataArea(Type,Area).

storeVerification(StoreID) :- activityFlow(SourceID,StoreID),
           compressActivity(SourceID,_).
storeVerification(StoreID) :- activityFlow(SourceID,StoreID),
           storeVerification(SourceID).
```

Transformations. Both the process model to specific domain model and process model to application model have to be modified. Indeed, the mapping source and target models changed, hence they have to be augmented with the new actions, activities and variables.

Results. It is now possible to use the small O&G editor to create processes, verify them and implement their behavior in a multi platform application. Figure 8 illustrates two processes modeled with an editor. The first one presents an error because there is no compression activity before the store activity. The second one is the corrected version of the first process which is validated by the expert system. Please, notice that figure 8 is an illustration of the process model. It is not produced with the Actinote 4.0 current editor.

This example shows that the addition of a domain specific in the platform induce modification in all the modules of the MDE approach. But, once these modifications have been made once, it is possible to generate as many different processes based on this specific domain asserting that they will be correct by construction. Moreover, as stated before, the editor created becomes easy enough to let the end-user model his processes himself.

5 Conclusion and Further Works

This paper presented that an established fact of the data in the O&G sector is that its interpretation relies heavily on human skill and experience: seismic data can be huge (up to hundreds of petabytes) and full of noise that needs to be manually cleaned. In order to justify the goal of the DataPipe platform, met by cooperating with a variety of specialists in a European project context: to alleviate work that would still be performed by human professionals. After a review of the different project stakeholders requirements, this paper presented Actimage model driven engineering approach to fulfill them. The paper then present an overview of the solution created to apply the approach: Actinote

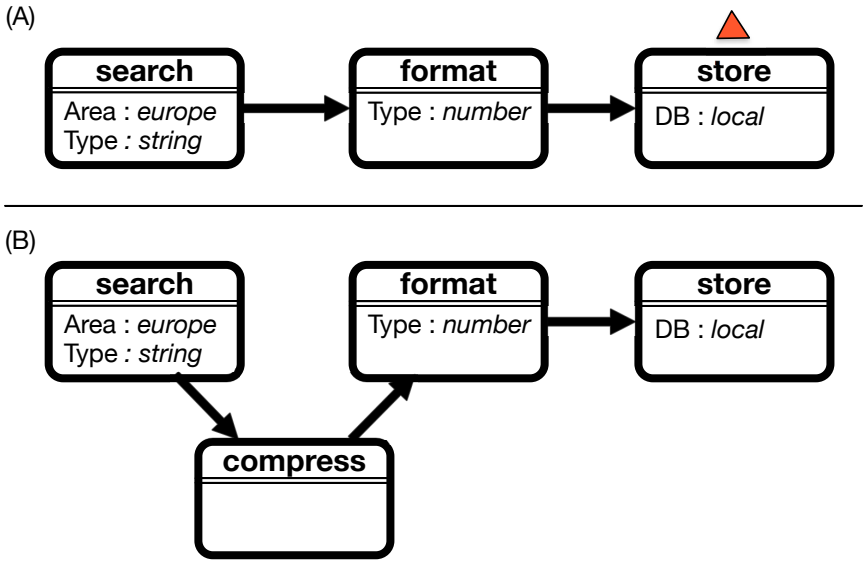

Fig. 8. small O&G processes models. (A) a model with a missing compress activity. (B) a validated model.

4.0 and how it is able to create, validate and implement model specific based data management processes.

Many products have been designed to solve the issues the Oil and gas industry is facing. The adaptivity of the product being a factor of the user acceptance, it seems therefore only clear that filling the gap between the users and the architect, as much as the gap between the noisy content and the normalized format, is essential. Heterogeneousness of formats having been a major subject of the oil and gas field for the past decades, it has been settled in the DataPipe project that giving control to industry specialists was the best approach to counter this environmental disparateness. The interactive Actinote 4.0 platform is the result of cloud-based engineering in that it uses adaptive behaviours to lower expectation differences between individuals and their devices. This brings a flexibility which can be perceived as a catalyst for the support of diverse digital intelligence media. Besides, the seismic stores of contents are arranged in a ubiquitous manner, hence an improvement of adjustability of data for both datamining and analysis purposes. The multiplatform aspects of the UI also mentioned in this article play an important role to the business logic adaptation one can observe using DataPipe software on mobile instruments or displays.

Datapipe project brought to Actimage knowledge in the expert system implementation and use. The size, context and complexity of the project proved to be a perfect opportunity to explore the MDE domain and apply it in, not only generic, but domain specific, user described, processes creation.

The interactions with the project partners allowed Actimage to acquire deep knowledge on specific domains such as the Oil & Gas data management, the print and the film industries. The collaboration in a multicultural context like Datapipes one brought some ideas and solutions that would never have rise otherwise.

Further works implies a deeper relation between the partners to enhance the current O&G knowledge database and then confirm the presented approach scalability to industrialisation. Another current experimentation is the implementation of other specific domain knowledge. Actimage currently works on a metrology based declination of Datapipe. And we also expect to test the two domains combination.

References

1. J. Bielak and D. Steeb, "Abstraction of multiple-format geological and geophysical data for oil and gas exploration and production analysis," Feb. 16 1999. US Patent 5,873,049.
2. R. D. Miller, J. H. Bradford, K. Holliger, and R. B. Latimer, eds., Advances in near-surface seismology and ground-penetrating radar. No. no. 15 in Geophysical developments series, Tulsa, Okla. : Washington, D.C. : Denver, Colo: Society of Exploration Geophysicists ; American Geophysical Union ; Environmental and Engineering Geophysical Society, 2010.
3. Ö. Yilmaz, S. M. Doherty, and Ö. Yilmaz, Seismic data analysis: processing, inversion, and interpretation of seismic data. No. no. 10 in Investigations in geophysics, Tulsa, OK: Society of Exploration Geophysicists, 2nd ed ed., 2001.
4. N. Krichene, "World crude oil and natural gas: a demand and supply model," Energy Economics, vol. 24, no. 6, pp. 557 – 576, 2002.
5. CDW, "High-performance computing in oil and gas," tech. rep., CDW, 2014.
6. HP, "Hp workstations for seismic interpretation," tech. rep., Hewlett-Packard Development Company, 2013.
7. M. Arenaz, J. Dominguez, and A. Crespo, "Democratization of hpc in the oil & gas industry through automatic parallelization with parallware," in 2015 Rice Oil&Gas HPC Workshop, 2015.
8. S. Mellor, A. Clark, and T. Futagami, "Model-driven development - Guest editor's introduction," IEEE Software, vol. 20, pp. 14–18, Sept. 2003.
9. Y. Lamo, X. Wang, F. Mantz, ø. Bech, A. Sandven, and A. Rutle, "DPF Workbench: a multi-level language workbench for MDE," Proceedings of the Estonian Academy of Sciences, vol. 62, no. 1, p. 3, 2013.
10. K. Czarnecki and S. Helsen, "Feature-based survey of model transformation approaches," IBM Systems Journal, vol. 45, no. 3, pp. 621–645, 2006.
11. A. Metzger, "A systematic look at model transformations," in Model-driven Software Development, pp. 19–33, Springer, 2005.
12. J. Hutchinson, J. Whittle, and M. Rouncefield, "Model-driven engineering practices in industry: Social, organizational and managerial factors that lead to success or failure," Science of Computer Programming, vol. 89, pp. 144–161, Sept. 2014.
13. M. Torchiano, F. Tomassetti, F. Ricca, A. Tiso, and G. Reggio, "Relevance, benefits, and problems of software modelling and model driven techniques—A survey in the Italian industry," Journal of Systems and Software, vol. 86, pp. 2110–2126, Aug. 2013.
14. OMG, "MDA Guide 1.0.1," 2003.
15. M. Voelter and S. Benz, eds., DSL engineering: designing, implementing and using domain-specific languages. Lexington, KY: CreateSpace Independent Publishing Platform, 2013.
16. WFMC, "Terminology and glossary," Tech. Rep. WFMC-TC-1011, Issue 3.0, Workflow Management Coalition, Feb. 1999.
17. C. Morley, Processus métiers et systèmes d'information: évaluation, modélisation, mise en oeuvre. Paris: Dunod, 2007.
18. W. Aalst, A. Barros, A. Hofstede, and B. Kiepuszewski, "Advanced workflow patterns," in Cooperative Information Systems (P. Scheuermann and O. Etzion, eds.), vol. 1901 of Lecture Notes in Computer Science, pp. 18–29, Springer Berlin Heidelberg, 2000.

19. W. Aalst, A. Hofstede, B. Kiepuszewski, , and A. Barros, “Workflow patterns,” Distributed and Parallel Databases, vol. 14, no. 1, pp. 5–51, 2003.
20. P. Jackson, Introduction to expert systems. Addison-Wesley Pub. Co.,Reading, MA, Jan. 1986.
21. A. Eardley and L. Uden, eds., Innovative knowledge management: concepts for organizational creativity and collaborative design. Hershey, PA: Information Science Reference, 2011.
22. F. Bourgeois, P. Studer, B. Thirion, and J.-M. Perronne, “A Domain Specific Platform for Engineering Well Founded Measurement Applications,” in Proceedings of the 10th International Conference on Software Engineering and Applications, pp. 309–318, 2015.

Document Management

Filippo Eros Pani[1], Beniamino Valcalda[2], Simona Ibba[1] and Simone Porru[1]

[1]Department of Electrical and Electronic Engineering, University of Cagliari,
Piazza d'Armi, 09123 Cagliari, Italy
[2]T Bridge, T Bridge S.p.A., Genova, Italy
{filippo.pani, simona.ibba, simone.porru}@diee.unica.it
b.valcalda@tbridge.it

Abstract. An analysis of one's organization is a necessary first step for any effective business strategy. Such an analysis needs to begin from one of the most notorious sources of costs and inefficiency: document management. This is an area where a transformation is in order to meet the need for an increase in company productivity and in process efficiency, while reducing operation costs. The proposed project is placed in the context of IT and Cloud Computing, and aims at creating an integrated system capable of providing a set of solutions and dedicated services for document management. This project is developed by the Department of Electrics and Electronics Engineering of the University of Cagliari and T Bridge S.p.A. It is financially supported by the Autonomous Region of Sardinia with European local development funds.

1 Introduction

Companies are undergoing a rapid transformation, under the pressure of an increasingly digital world and increasingly competitive markets.

Rationalization, simplification and automation of processes have become an essential leverage to free resources, which in turn can be invested in technology to accommodate business innovation and generate competitive advantage.

Public Administration in Italy has an established legal framework defining the digitalization of the administrative action. The latest reforms focused on new technology as the main tool to interact with citizens, with a consequent impact on communication processes between public institutions and private citizens, as well as on the internal organization and instruments of public institutions themselves.

This paper is structured as follows: the second section describes the context in which the proposed project lies, while the section following it describes activities and objectives in detail, and the schedule of the project is outlined. The last section hosts our final observations about the project.

2 Context of Research Proposal

The potential market segment to which document management is associated offers a swath of products and services capable of supporting communication and

interoperability among different subjects. These products and services target all business users, such as institutions and public administration.

Regarding information flows and communications, business users who choose the proposed solution will be potentially able to save both time and money.

Companies, under the pressure of an increasingly digital world and competitive markets, are adapting at a fast pace. Processes rationalization, simplification and automation have become a fundamental means for unleashing resources to invest in the technology necessary to foster business innovation and create competitive advantage. In this scenario in order to develop a successful strategy, managers must first analyze the organization they work for, starting from the information management, being one of the most common sources of inefficiency; then, also, the management of all of the processes related to documents. This is the aspect where a deep transformation is required, as it reflects the need to increase business productivity and process efficiency while reducing the operating costs.

Public administrations and specifically the local ones, organizations usually relatively complex, could especially benefit from these solutions: an increased document management capability without the need for conspicuous investments to increase disk space, application servers, front-office extensions and back-office maintenance.

To provide a effective and tangible response to each and every one of these challenges, it becomes necessary to make the most simple choices, which are also easily shareable and implementable in a short time: the solution is the employment of a series of services provided by a website, where the users will be able to fulfill all of their needs: tasks organization; communication; electronic documents creation, digital signing, marking, registering, safe submission with return receipt on multiple channels (mail, email, fax, certified email, SMS), and storage in accordance with the law.

By leveraging these services, companies and professionals will be able to organize their work without worrying about all the aspects related to the management of paper documents (printing, mailing, cataloguing and storage), while a public institution will benefit from being delivered from all of the back-office processes requested by the management of the electronic document in different departments (registering, general affairs, accounting, and so on).

3 Research Project Description

This project aims to create an innovative software platform for document management. The platform will be offered to users “as-a-service”, and be used to manage communication, electronic document creation, digital signature, digital registration, document recording protocols, and secure transmission. It will ensure messages will be received through several channels (mail, email, fax, certified email, text messages, voice), and that archiving and storage will be performed according to local laws.

The main goal is to provide end users and companies (that already use Gmail, Google Calendar, and Google Docs) with a suite of web-based RIA tools both safe and reliable, which will serve as a means of communication with other institutions, such as public ones, but also banks, companies, freelancers, etc. The applications dedicated to

document management will be created leveraging the open source development toolkits provided by Google.

An advanced Lean-type approach to software development is deemed to be suitable for the project. The Lean approach was meant to be used to support systems' maintenance and evolution, and the first applications of Kanban actually concern maintenance processes [1]. Moreover, it allows to effectively manage the development of several concurrent projects by a single team, that is not an ideal situation but a very common one. It is indeed a fairly flexible approach, which meets the actual needs of the organizations that must develop and maintain software, especially if compared to other less recent approaches.

The Lean approach is spreading fast. In order to avoid past mistakes, however, it is vital for the "Lean" approach to undergo a rigorous and experimental observation [2] [3]. Given the radical difference of this approach in comparison to more common models of software development organization, it will be important to understand the impact of its introduction at marketing, organizational, and company level [4], [5].

The solution we intend to create in this project makes use of powerful and widely tested development tools, already adopted by wide developer community, which can provide support to implementers.

As for the reference market, this solution will be able to be transmitted through the widespread usage of the tools provided by Google.

Concerning the production model we adopted, the development tools we used ensure a high productivity in the solution selection process, guaranteed by software development platforms provided by Google.

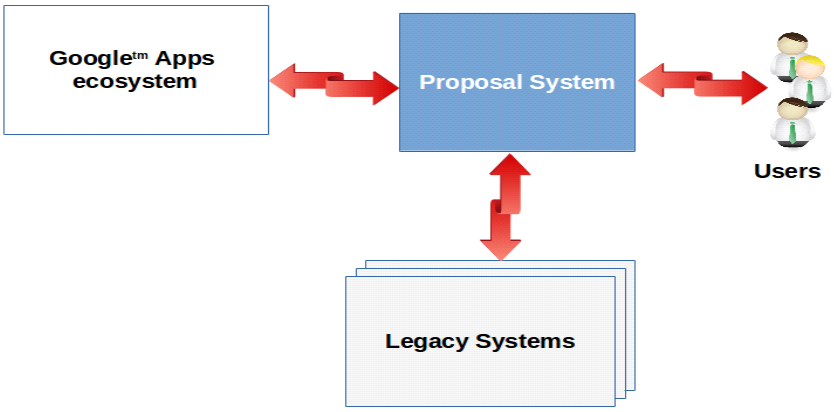

Fig. 1. Proposal system and interactions.

With regard to the prototype, modularity will be a regarded as a key feature of the architectural solution chosen for the system, since it requires the integration of software components, which are themselves modular, interoperable and flexible, thanks to the adoption of open standards.

Attention will be paid on the building of a prototype that solves the complexities arising from the interaction of separate, although interdependent, processes. An interoperability infrastructure between solutions and specific services will be studied, to fulfill each user request through an integrated and transparent manner, and also taking the Cloud Computing paradigm into account [6]. The paradigm offers several

advantages in terms of reliability, scalability, resource optimization, cost reduction, failure recovery, environmental impact, etc. [7].

The project is highly innovative and ensures acquisition of valuable new knowledge for all the involved parties, especially because of the dense cooperation network it will foster. Knowledge and technology dissemination will be ensured by the intrinsic features of the project, being it an open-source project: it is open to external contributions, particularly from those entities that see the project as a solution for their business challenges. The open-source diffusion lead us to face a revolution in the IT market,which in the recent years made us witness the evolution of the business model from the development of stand-alone software packages to the so-called "System Integration", thus proving the shifting of the investments on services.

New knowledge acquisition will be ensured by the tight collaboration between T Bridge and University of Cagliari, together with the dense cooperation network which will be generated, also with the aim of creating clusters of IT companies.

The project will lead to a considerable advancement in the state of the art, as it is concerned with the process management related to Service Oriented Architectures (SOA). It will foster the integration with the tools for the management of the documents involved in such processes, and the management of all the data deriving from those processes [8] [9].

3.1 Project Subdivision

The project officially began on October 1, 2015, and its conclusion is estimated to be on September 28, 2018.

The project covers a number of operational stages. Every stage of the working plan is organized in Work Packages (WP), parallel phases in which operational objects are reached with work group activities, through the production of expected results and products and the application of a specific methodology.

The WP included in the project are five:

- WP 1: State of the art evaluation.
- WP 2: Software development tools and technologies assessment.
- WP 3: Applications and services definition.
- WP 4: Architectural solution design.
- WP 5: Prototype development.

Below is a brief description of each phase of the project development.

3.1.1 WP 1 - State of the Art Evaluation

The first work package is focused on the study of the wide-ranging needs of innovative Cloud Applications, with the aim of performing a proper selection, and drive the project towards the development of the most disruptive functionalities. These functionalities will be selected based on their potential practical impact on the industrial project.

3.1.2 WP 2 - Software Development Tools and Technologies Assessment

This work package concerns on the one hand the design of an innovative Lean approach for supporting software development in a Cloud environment, and possibly in a distributed and collaborative context, on the other hand the design of an integrated Lean approach oriented towards the management of the business processes of the partner company. The activities will go beyond the study and design of a Cloud development methodology, as also the idea of "developing software for the Cloud leveraging the opportunities offered by the Cloud" are to be followed to determine which key modules to select in a stand-alone environment.

3.1.3 WP 3 - Applications and Services Definition

Challenges specifically Cloud-related will be here studied, together with the effectiveness of the solutions currently proposed on the market, in view of future developments in the medium to long term. The tasks in this package have a primary role in the project as, currently, there are no similar studies that use data from student projects, and that could be regarded as representative of an industrial or professional development.

3.1.4 Wp 4 - Architectural Solution Design

The aim of this package is to study suitable architecture solutions, and finally design it. Basically, the final objective is to analyze a possible technological infrastructure, which can integrate the primary issues and those related to the associated applications.

Cloud computing will be taken into account from the point of view of technological cooperation, and of cooperation logics between applications and data, which can come from both different systems and different entities.

The infrastructure will be able to handle document management streams in support of an effective application engine for applications management [10], [11], in addition to a presentation through Google apps tools, in a SOA logic. The ultimate goal will be the definition of an easy-to-use tool for the end user.

3.1.5 Wp 5 - Prototype Development

In this phase, the prototypes for all the software components of the document management system will be implemented and validated. The development will require specific modules, that will consider key features such as modularity, interoperability, and flexibility, by leveraging open standards. Special attention will be paid on the development of a prototype which, in the context of applications, will solve all the issues which originates when making distinct yet interdependent processes interact.

Another aspect that will be taken into consideration in the prototype development phase is the SOA integrated security management for the Cloud, in a dynamic implementation, with focus on access authorizations. Moreover, in this work package, the results of the project will be published, through specific planning activities and the organization of workshops, participation to international conferences, and scientific articles.

3.2 Project Results Protection and Valorization

All the acquired knowledge and expertise will be protected through the ownership of the know-how in the specific context in which the project will be developed. As the industrial partner features an open source business model, and most of the tools selected for the system are also open source, it is not wise to perform expensive procedures for patenting all the proposed methodologies and practices. Also, this would not be suitable because obtaining and, particularly, making patents respected in such an intangible context is extremely difficult. Moreover, it should be noted that, as of today, software in Europe can not be patented, considering all the currently available procedures for the patenting of both methodologies and practices.

Enterprise protection is a process based on professionalism: its foundations are its superior know-how and the know-how continuous development with respect to the competitors. This will make the company rely exclusively on it acquired expertise.

Enterprise value will be obtained through added value services sellings (services development, installation, training, customization, etc.), which will be all the more marketable and remunerative the greater the specific know-how and visibility of the company.

Scientific and technological results which will be obtained through research are:

- the study of the interoperability infrastructure in its basic applicative challenges;
- SOA integrated security management for Cloud support, paying special attention on access authorizations;
- the application engine for the applications management and a presentation via Google Apps tools;
- the definition of an innovative architecture for integrating a service development tools suit;
- the industrial results which will be obtained through research are:
- the acquisition of the relevant know-how related to open contexts and their application on the service-oriented Web, in order to increase users productivity, and quality;
- the development of specific modules which take into account all the fundamental aspects, such as modularity, interoperability, and flexibility, through the use of open standards;
- the acquisition of the know-how concerning Cloud processes, in a model where the application-centered aspect is integrated with a document-centered approach;
- the acquisition of the relevant know-how related to the use and integration of open-source tools for supporting software agile development.

4 Conclusion

The solution will be an easy-to-use tool for end users, and, as it is based on Google technology, will offer high security with low downtime levels. The results of the project will be disseminated through workshops, conference participation and international scientific publications.

The presented project financed by the Autonomous Region of Sardinia with European funds (Single Programming Document 2007-2013 - P.O. FESR 2007-2013 -

Line of Activity 6.2.2.d - Interventions to support competitiveness and innovation, under the Regional Committee Resolution no. 33/41 of 08/08/2013).

The creation of the platform will stem from the strategic partnership between T Bridge S.p.A. and the Department of Electrical and Electronic Engineering (DIEE) of the University of Cagliari. The purpose of this choice to use the results of fundamental research as well as of industrial research to elaborate an innovative prototype.

This project is coherent with the strategic objective of the regional planning in Sardinia, since it aims to implement innovative methods of the ICT sector in the library industry, and it complies with the objectives described in the Regional Strategic Document (Documento Strategico Regionale, DSR) 2007-2013 for Sardinia.

Acknowledgements. Simona Ibba and Simone Porru gratefully acknowledges Sardinia Regional Government for the financial support of his PhD scholarship (P.O.R. Sardegna F.S.E. Operational Programme of the Autonomous Region of Sardinia, European Social Fund 2007-2013 Axis IV Human Resources, Objective l.3, Line of Activity l.3.1).

References

1. Anderson, David J.: Lessons in Agile Management: On the Road to Kanban. Blue Hole Press (2012).
2. Corona E., Pani F. E.: A Review of Lean-Kanban Approaches in the Software Development. In: Transactions on Information Science and Applications, Issue 1, Vol. 10 (20139, Print ISSN: 1790-0832, E-ISSN: 2224-3402.
3. Leffingwell, D.: Agile software requirements: lean requirements practices for teams, programs, and the enterprise. Addison-Wesley Professional (2010).
4. Buschettu, A., Sanna, D., Concas, G., Pani, F. E.: A Platform based on Kanban to Build Taxonomies and Folksonomies for DMS and CSS. In: Journal of convergence, Vol.6, No.1 (2015), pp. 1-8.
5. Buschettu, A., Concas, G., Pani, F. E., Sanna, D.: A Kanban-based Methodology to Define Taxonomies and Folksonomies in KMS. In: Park, J. J. J. H., Stojmenovic, I., Jeong, H. Y., Yi, G. (Eds.): Computer Science and its Applications Ubiquitous Information Technologies. Lecture Notes in Electrical Engineering, Springer Berlin Heidelberg (2015), pp. 539-544.
6. Lewis, G. A.: The Role of Standards in Cloud-Computing Interoperability. Software Engineering Institute, Carnegie Mellon (2012).
7. 10 Workday, Inc.,: Critical Requirements for Cloud Applications: How to Recognize Cloud Providers and Applications that Deliver Real Value (2012)
8. Sharma, R., Sood, M, Sharma D.: Modeling Cloud SaaS with SOA and MDA. In: Advances in Computing and Communications - Communications in Computer and Information Science (2011).
9. Pani, F. E., Concas, G., Porru, S.: An Approach to Multimedia Content Management. In: Proceedings of 6th International Conference on Knowledge Engineering and Ontology Development (2014), pp. 264-271.
10. Boese, S., Reiners, T., Wood, L. C.: Semantic Document Networks to Support Concept Retrieval. In: Encyclopedia of Business Analytics and Optimization. Hershey, IGI Global (2014).
11. Luo, J., Meng, B., Tu, X., Liu M.: Concept-based document models using explicit semantic analysis. Granular Computing (GrC), In: IEEE International Conference on. (2012), pp. 338-342.

Author Index