NRG4CAST
FP7-2012-NMP-ENV-ENERGY-ICT-EeB
Contract no.: 600074
www.NRG4CAST.org

NRG4CAST

Deliverable D8.3

Best Practices guide

Editor:	Andreas Androutsopoulos, George Markogiannakis, CRES
Author(s):	Tatsiana Hubina, Diego Sanmartino, CSI; Klemen Kendra, Maja Skrjanc, JSI; Yannis Chamodrakas, Kostas Kalaboukas, SLG; George Markogiannakis, Andreas Androutsopoulos, CRES; Steffen Nienke, FIR, Irene Koronaki, NTUA; Simon Mokorel, ENV; Caterina Calefato, IREN
Reviewers:	Klemen Kendra, JSI; Steffen Nienke, FIR
Deliverable Nature:	Report (R)
Dissemination Level: (Confidentiality)[1]	Public (PU)
Contractual Delivery Date:	May 2015
Actual Delivery Date:	May 2015
Suggested Readers:	ICT software developers, software companies
Version:	1.0
Keywords:	NRG4CAST components implementation problems – solutions, lessons learned

[1] Please indicate the dissemination level using one of the following codes:
• **PU** = Public • **PP** = Restricted to other programme participants (including the Commission Services) • **RE** = Restricted to a group specified by the consortium (including the Commission Services) • **CO** = Confidential, only for members of the consortium (including the Commission Services) • **Restreint UE** = Classified with the classification level "Restreint UE" according to Commission Decision 2001/844 and amendments • **Confidentiel UE** = Classified with the mention of the classification level "Confidentiel UE" according to Commission Decision 2001/844 and amendments • **Secret UE** = Classified with the mention of the classification level "Secret UE" according to Commission Decision 2001/844 and amendments

Executive Summary

The Best Practices Guide intends to provide information regarding the experience gained through the different stages of the NRG4CASTproject and the preparation of the NRG4CAST tool to a variety of pilot cases. The developed tool is the outcome of a variety of different software and computational processes which came together and linked accordingly so as to provide the NRG4CAST tool features. The completion of the tool task was accompanied by numerous obstacles and problems that had to be solved. In this guide, the obstacles discovered, the solutions found as well as the lessons learnt are presented in a prescribed format. The main layout of the guide pinpoints the important aspects (components and sub-layers) of the project, explains their concept, objectives, the methodology they followed and lists the problems they were faced with and the solutions found, together with the knowledge gained. The experience obtained is being presented in a concentrated form, but also in detailed according to the sub-components way. So, the guide presents the core components of the tool: Data collection, Analytics, and Visualization following the above layout concept mentioning also all the important sub-fields that have been dealt with during the project.

In addition, the problems faced during the implementation of the developed tool to the pilot cases together with the lessons learnt and recommendations derived have also been recorded and presented.

Finally, some future research areas towards the technical parts which have been identified through the project duration are stated together with some ideas for additional features of the NRG4CAST tool.

Table of Contents

List of Figures

Abbreviations

CAS - Central Authentication Service

CEP - Complex Event Processing

ETL - Extract Transform Load

GAV - Global As View

LAV - Local As View approach

NLP - Natual Language Processing

OGSA-DAI - Open Grid Services Architecture Data Access and Integration platform

PMC - Prediction Manager Component

SSO - Single Sign On

1 Introduction

This report presents the experience gained and the lessons learnt throughout the implementation of the "NRG4CAST – Energy Forecasting" project. It includes the methodology, the structure, the analytics, the data mining, extraction, cleaning as well as visualization techniques followed during the development and implementation of the NRG4CAST tool. Focus was given in the ICT features of the NRG4CAST tool, the different software-origin interaction, the layout of the developed tool, and the issues that arose during the preparation of the tool, its implementation phase, and the solutions found. Additionally, the guide describes the experience came out from the collaboration of the various origin project partners, provides features, special points of interest, experience gained from the run of the pilot cases from the prototype tool, and provides with ideas / thoughts regarding further research areas on the fields that have been dealt with during the project.

The current report utilized the information and results of the project purely technical work packages that constitute to the research activities of the project and have advanced to a high level of completeness. It includes the work carried out for Data processing – cleaning - modelling, Real time monitoring, alerting and reasoning, as well as the integration procedures followed for the toolkit creation (see Figure 1).

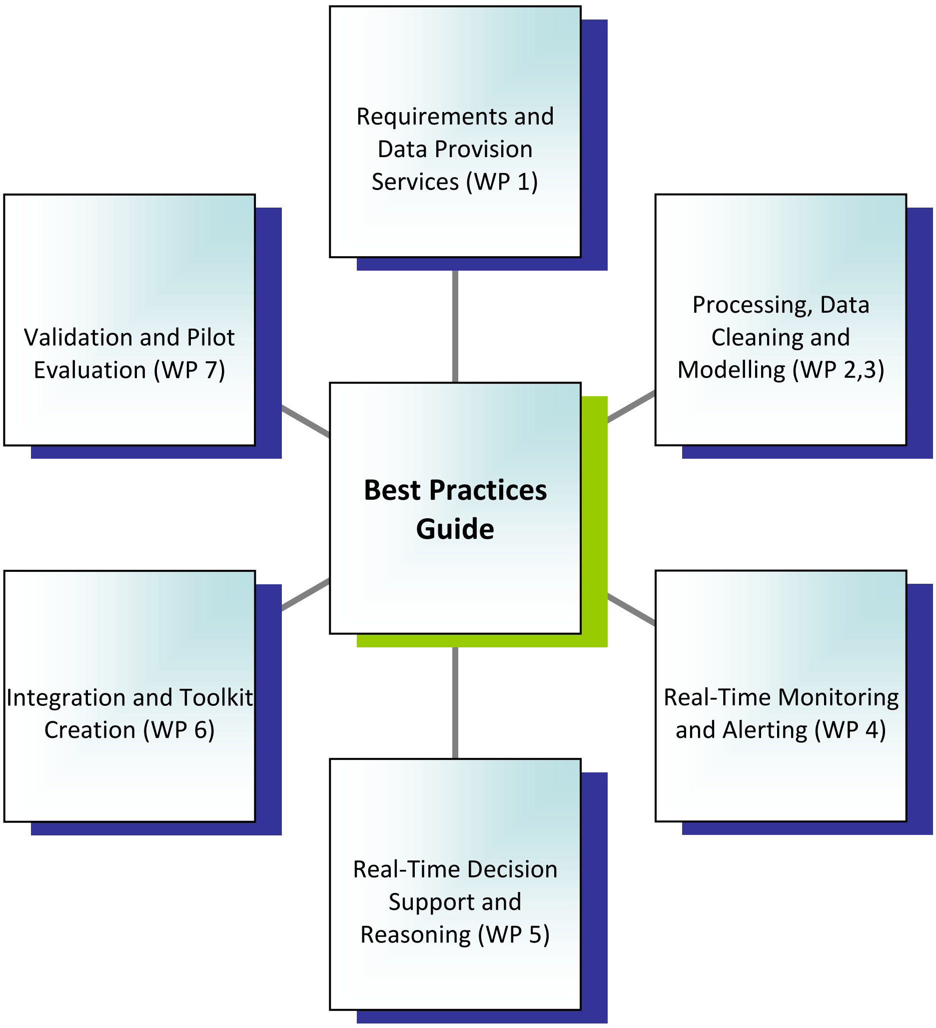

Figure 1: Input of different processes to the Best Practices guide.

The guide follows a common layout on the important aspects of the project (technical WPs) which includes the objectives of the individual NRG4CAST components, the philosophy used, the approach they followed for the implementation, the issues they faced and the solutions found, the lessons learned from the integration procedure, the perspectives that came out of this work, and the research areas that emerged during the project duration. The overall approach utilised in this report is presented in Figure 2.

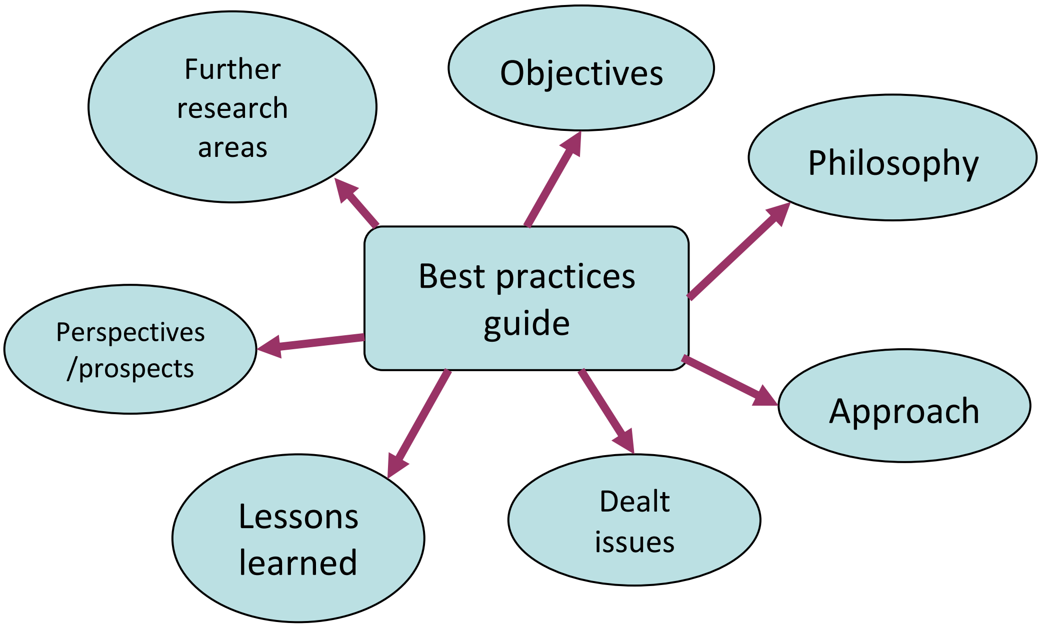

Figure 2: Aspects of the Best Practices Guide.

Regarding structure, the Best Practices Guide provides with a section regarding the experience gained out of the development of the NRG4CAST tool during the project lifetime (Ch.2). The NRG4CAST solution functionalities are briefly presented in Chapter 3, together with its infrastructure.

The main NRG4CAST components are described in the next Chapter, namely Data collection, Analytics, and Visualization as well as authentication and authorization functions in view to the experience gained from their implementation.

The experience gained from the implementation of the Prototype NRG4CAST on the Pilot cases is presented in Chapter 5 together with the problems faced, the solutions found and recommendations. Finally, some concentrated findings regarding future research areas of interest can be found in Chapter 6.

2 Main experience gained

The development of the NRG4CAST platform and the work out of the specific pilot cases, resulted in a couple of achievements and good starting points for others, dealing with the same kind of platform. Technologically, these are the main achievements of the project.

- **Data gathering infrastructure**
  Practice from many different EU funded project has shown before, that having reliable access to all the data is essential for the smooth development of the project tasks. Such an infrastructure has been developed at the bottom tier of the NRG4CAST platform. Data availability was – with such a service – no more in a domain of the pilot partners. Additional degree of reliability has been added to the system. Upper layers have always been able to access the data or even – demand a full reload (for example Prediction Manager) if needed.

  Additional value has been added to the system with the dynamic registry of sensors, employing many different data adapters for many different data sources. It is impossible to make such a component completely dynamic, but a great deal of data sources is supported and can be now solved in a matter of a few clicks.

- **Using Kalman filtering for data cleaning**
  Kalman filter is a method that has been known to the data mining community for many years. It is a widely spread algorithm that can be applied to many different domains. In NRG4CAST, we use it as a short time predictor and as an error estimator. With this knowledge, in NRG4CAST, a dynamic data cleaning environment for detection of single outliers has been developed. The environment should be expanded with more classical approaches to reach its full potential for accuracy and recall, but the method itself is different and innovative.

- **Prediction Manager – full multi-modal data fusion**
  Machine learning methods and its applications have been well researched in the previous decades. Moving the setup into a streaming environment removes a decade or two of experience, but it is still a domain, that has been investigated quite well. The achievement of the NRG4CAST platform is that it is able to fuse multi-modal data (different kinds of sensors with different time granularity and delays, different kinds of metadata – including data deduced from purely textual sources, different kind of forecasts). The task, that might seem trivial at first, brings a lot of different challenges into the picture and the results of the NRG4CAST might serve as a very good guide for developers of similar applications.

- **Abstractions for visualizations**
  Basic functionality of sensor system enables the user to visualize a time series, belonging to a certain sensor. With NRG4CAST, the user (or developer) has an opportunity to visualize also different aggregates of sensor measurements or even composites, based on sensor data (different mathematical operations, basic histogram features, etc.).

- **Dynamic configuration of visualisation component**
  The visualization part of the NRG4CAST system is characterized by dynamic configuration. This flexible solution has allowed for implementation of a number of pilots and reports starting from a single application and to manage the layer of visualization and reporting in a more effective way. It's possible to easily add and remove reports and functionalities within the prototype. This solution can be definitely utilized for other visualization solutions to be developed in the future.

- **Visualisation component to be integrated into smart data platform**
  On behalf of Piedmont Region, CSI develops the smartdatanet, available at: http://www.smartdatanet.it/. This platform integrates various types of data coming from Internet

of Things and Internet of People. From a multidisciplinary prospective, these data can be used for implementation of new applications. Based on the smartdatanet, CSI has developed Yucca platform: http://www.smartdatanet.it/yucca-platform which gives a possibility to create an application, to interconnect applications, social network and data in general. Moreover, Yucca platform provides options for data gathering and for advanced data analysis, creation of new applications and solutions.

The idea is to integrate NRG4CAST system within the smart data platform. Working on the Energy efficiency in public owned buildings, mainly public offices of Turin business model, the data sources and data integration process will be modified, however, the visualization part will be re-used completely. The experience gained through the development of the visualisation and the report managing part of the NRG4CAST systems resulted in a "ready to use" dashboard to be integrated within the smart data platform.

- **Best practices and tools**
  There is a number of remarkable experiences and tools, developed during the project, to be utilised in other contexts and applications. For example, an advanced search option has been developed, which gives user a possibility to easily navigate within the sensors hierarchy and to locate the sensors of interest.

Many small innovations or useful pieces of protocols or experience have been gained through the project and they are described in more detail in the following sections.

3 NRG4CAST

In this chapter, the functionalities of the NRG4CAST tool are briefly presented together with the NRG4CAST infrastructure so as to form a complete picture of the different aspects of the tool architecture, their interaction and the possible problems that can arose during their implementation.

3.1 NRG4Cast main functionalities

NRG4Cast project's scope is to develop advanced solutions for predicting behaviour of local energy networks for the three fundamental scenarios:

- Predicting energy demand on several network granularity levels (region, municipality, city, business, household and energy service provider);
- Predicting energy network failures on interlinked local network topologies;
- Detecting short term trends in energy prices and long term trends in national and local energy policies.

Through the project duration, the NRG4Cast tool developed and set out real-time monitoring and prediction services based on machine learning, trend detection, predictive analytics, optimisation and reasoning capabilities, in order to provide a unique service for energy planning, network failure and energy price prediction. The NRG4CAST tool has the ability to

- assess different types of information, combined from internal and external sources, i.e. historical and current information from the local energy distribution network (energy consumption and network devices status), information about the local environment (natural and social), energy prices (national and international), etc.
- cope with highly dynamic environment of energy distribution networks including emerging renewable energy sources, dynamic network topologies with electric vehicles, local energy storage, virtual plants and others.
- be applied in any type of energy distribution network (electricity, gas, heating etc.), and deal with mass amount of multimodal data in real-time.

In order to achieve its objectives, the concept of NRG4CAST toolkit architecture is to integrate a set of real-time energy monitoring and prediction software components so that it will enable their provision as services with specific APIs to external systems and applications. A crucial requirement for the specification of the toolkit architecture was the ability to combine different types of information provided by different information systems and databases. According to the above and taking into account the requirements of the pilot cases, the NRG4CAST toolkit architecture follows a multi-tiered Web Service Architecture approach. The layout of the toolkit architecture is depicted in Figure 3.

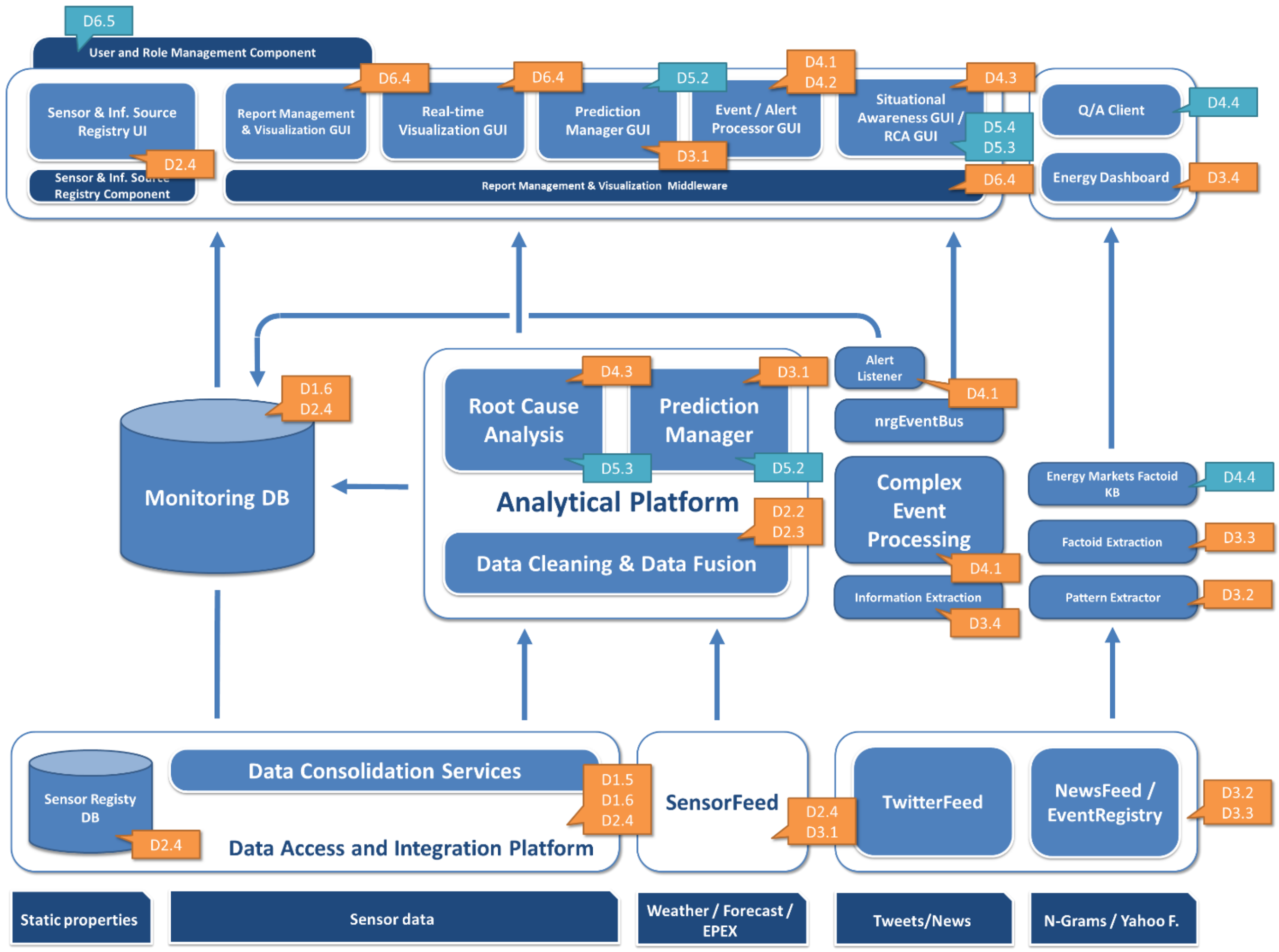

Figure 3: NRG4CAST architectural diagram.

3.2 NRG4CAST infrastructure

In an average deployment scenario, the NRG4CAST toolkit will integrate hundreds of sensors and other information sources sending data every hour or even more frequently. Such an average scenario is translated into hundreds of millions of rows in the Monitoring Database which could be further increased into several billions if we envision a continuous use of the NRG4CAST toolkit for a number of consecutive years. At the same time, the execution of online data cleaning, aggregation, forecasting and other analytics algorithms on a very high volume of data in the Analytical Platform poses also the requirement for increased processing power and memory.

Therefore, an average deployment of the NRG4CAST toolkit necessitates the installation of the NRG4CAST toolkit in 3 server machines:

- **Database Server** (Monitoring Database, Sensor Registry, Data Access & Integration Platform)
  - Recommended specs for up to **X** GB Monitoring database:
    - **RAM**: 4 GB OS / **X** GB DB
    - **HDD:** 11 * **X** GB space / SSD - RAID5
    - **CPU:** 2 Eight Core Processors (e.g. Intel Xeon E7)
- **Analytics Server** (EventRegistry, Analytical Platform, Complex Event Processing)
  - Recommended specs:
    - **RAM**: 4 GB OS / 32 GB Analytics / 92 GB Event Registry

        - **HDD:** 3*2 TB space
        - **CPU:** 2 Eight Core Processors (e.g. Intel Xeon E7)

- **Web Server** (UI Components)
    - Recommended specs:
        - **RAM**: 4 GB OS
        - **HDD:** 500 GB space / SSD
        - **CPU:** 1 Eight Core Processor (e.g. Intel Xeon E7)

The deployment of the NRG4CAST toolkit may be done either in-house or through purchasing Infrastructure-as-a-Service (IaaS), i.e. in the “cloud”. The decision should be based on the desired balance between, on the one hand, cost and, on the other hand, data confidentiality and management control criteria.

4 NRG4CAST components

This chapter presents the NRG4CAST components, Data collection, Analytics, and Visualization, as well as authentication and authorization functions, the main challenges faced during their implementation to the tool, and how they were overcome. Additionally, the Authentication and Authorization functions are also presented here.

4.1 Data collection

4.1.1 Data access and integration platform

4.1.1.1 Concept

Data Integration is the process of combining data coming from heterogeneous sources in order to provide a unique query interface to the systems and/or the users requiring access to the data. The main approaches for data integration are [1]:

- The **data warehouse approach** according to which data are extracted from the separate data sources, are subsequently transformed and, in the final step, are loaded into a data warehouse with a single schema. This approach is usually referred to as ETL (Extract-Transform-Load) and it is characterized as tightly coupled.
- The **mediated schema approach** according to which data are retrieved directly from their original databases and only the query interface is provided in a unified way. The unified query interface necessitates a mapping between the original and the mediated schemas: in the "Global As View" (GAV) approach, through a mapping from entities in the mediated schema to entities in the original sources, in the "Local As View" (LAV) approach through a mapping from entities in the original sources to the mediated schema. This approach is characterized as loosely coupled and it is more feasible for datasets that are frequently updated since there is no need for explicit synchronization.

In the context of the NRG4CAST project a software platform has been developed in order to integrate data coming from hundreds of different sensors and other sources as well as necessary metadata.

4.1.1.2 Problems faced and solutions found

The main issues faced during the data integration process are the following:

- To **design the schema for the metadata database** in order to flexibly adjust to the requirements of different pilot cases. This issue was resolved through the design of a hierarchy of NRG4CAST entities. This hierarchy consists of: a) N levels of energy consumption (or production) centers; b) energy managed NRG4CAST objects which can be further instantiated to buildings, lamps, charging stations etc. which belong to one or more consumption centres; c) sensor nodes which are situated in the NRG4CAST objects; and d) sensors of sensor types that can be parameterized. The metadata schema has been proven to work with the divergent requirements of the different pilot cases.
- To integrate **real-time data** coming from hundreds of sensors. To resolve this problem, initially we considered the use of the mediated schema approach and therefore we adopted the Open Grid Services Architecture Data Access and Integration platform (OGSA-DAI) [2]. However, due to the need to provide access to aggregated values of cleaned sensor data and not to raw data we finally opted for the Data Warehouse approach through the design and development of the Monitoring Database [3], [4] which joins sensor aggregates, predictions and alarms with the metadata. The proper function of such an approach necessitated the implementation of real-time streaming mechanism continuously streaming sensor data to the Data Cleaning & Fusion component as well

as the provision of several APIs for pushing and pulling sensor data aggregates and predictions and a WebSocket listener for the storage of alarm signals from and to the Monitoring database [5].

- To support a **variety of data interfaces** (FTP, HTTP, SOAP, REST, Relational DBs) of different data sources which moreover have different modality (Relational Result Sets / CSV files/ JSON strings/ SOAP XML), format and structure. This challenge has been faced through the implementation of relevant functionalities in the Sensor Registry and Consumption Centre Configuration Component which is presented in section 4.1.2.

Another important problem that arose during the performance testing of the platform has been the high data retrieval latency due to the big volume of the data requested (e.g. sensor aggregate values spanning 2 years). A high latency can make the use of the system dysfunctional and therefore it posed a big challenge for the NRG4CAST software developers. In order to resolve this issue 2 independent measures have been taken: a) the first one has been to cache big volumes of frequently requested data at the front-end of the NRG4CAST toolkit; b) the second one has been to plan and perform the migration of the Monitoring Database to a dedicated server machine with high computing power, sufficient memory, and high-end hard disk drives. The first measure has been already implemented and has led to a significant improvement of the NRG4CAST performance whereas the second measure is still under progress and its results in relation to the data retrieval latency will be reported appropriately.

4.1.2 Sensor Registry and Consumption Centre Configuration Component

4.1.2.1 Concept

Sensor Registry and Consumption Centre Configuration is a part of the deployment and installation of a specific instance of the NRG4CAST system. Specifically, it is related both to the initial release and installation of a software system as well as the adaptation and extension of a software system that has been previously installed [6]. As such, the specific process is responsible for the creation and the configuration of all necessary resources that is the hierarchy of the consumption centers, the NRG4CAST managed objects, sensors and nodes. Therefore, Sensor Registry and Consumption Centre Configuration provides a Graphical User Interface and wizards for the configuration of all relevant metadata, data sources connection methods and details as well as for mapping the sensor data attributes to the schema of the Monitoring database and the inclusion of new sensor data sources in the streaming process from the Data Access & Integration platform to the QMiner analytical platform.

4.1.2.2 Problems faced and solutions found

The main issue faced in relation to the Sensor Registry and Consumption Centre Configuration process has been the need to configure the data fetching mechanism for each respective sensor according to a plethora of different data provision interfaces (FTP, HTTP, SOAP, REST, Relational databases). In addition, retrieved data have different modalities (Relational Result Sets / CSV files/ JSON strings/ SOAP XML), format (e.g. different data formats) and structure (different number and types of data attributes). In order to resolve this issue apart from creating the appropriate graphical wizards, the Sensor Registry database was extended with tables holding configuration information which can be dynamically reconfigured reconfigured (see Figure 4).

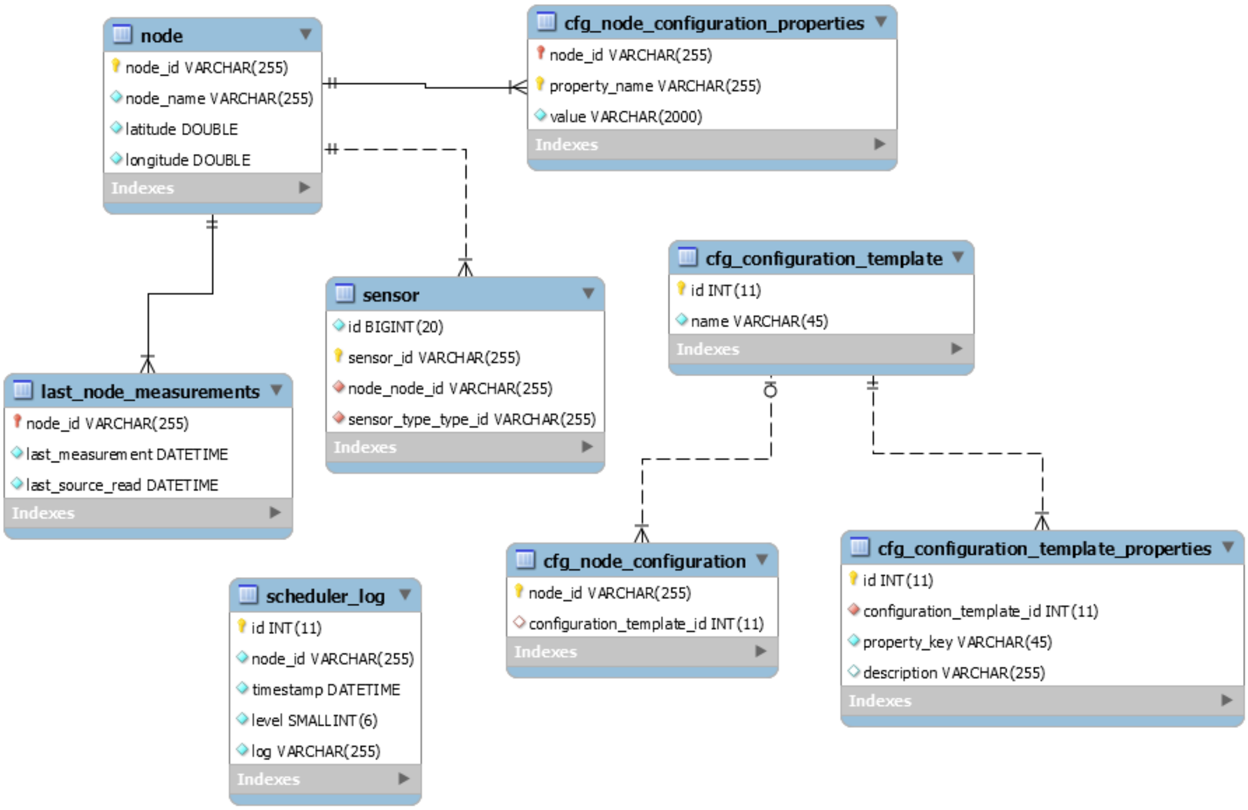

Figure 4: Configuration template tables.

4.1.3 Data Cleaning

4.1.3.1 Concept

Within the data cleaning prototype we have addressed an issue of detecting single false measurements (outliers) in a slowly changing stream of sensor measurements. Data cleaning is being done very early in the whole data distribution pipeline, therefore it demands effective yet non-complex (in the sense of data fusion) approaches.

The main approach utilizes a simple auto-regressive prediction method (2nd degree model for Kalman filter has been selected) with a tolerance gap for the expected interval of the next measurement. If a measurement falls outside of the gap, then the algorithm assumes it is a false one. The principle is depicted in Figure 5.

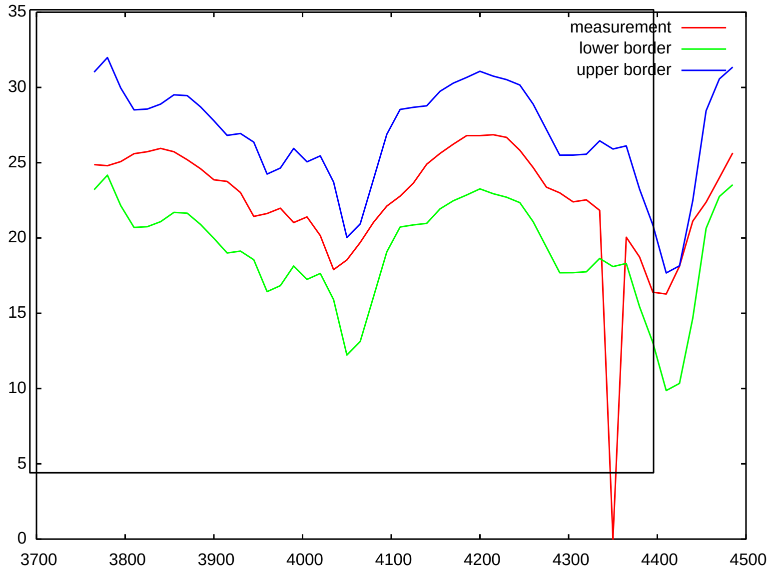

Figure 5: Illustration of identifying outliers with a short term autoregressive predictive method.

4.1.3.2 Problems faced and solutions found: Instability problem

Instability problem was the main problem we faced during the implementation and testing of Data Cleaning. It has been described in D2.3 [1] in section 2.4.3. The principle has been proven to be unstable and the Kalman filter prediction had diverged in some cases. If the filter encounters a false negative it relies on the prediction model, which can then move the lower/upper boundary so that no measurement ever again fits the criteria (see Figure 6).

A possible workaround includes an artificial increase of *a posteriori* variance (see band enlargement after false negative in the Figure 7). Theoretically, changing variance is mathematically incorrect, but in practice it has proven to be efficient. Besides, variance in the case of slowly changing values of sensor data soon converges to the vicinity of the correct values. Further research on the proposed approach is needed [8].

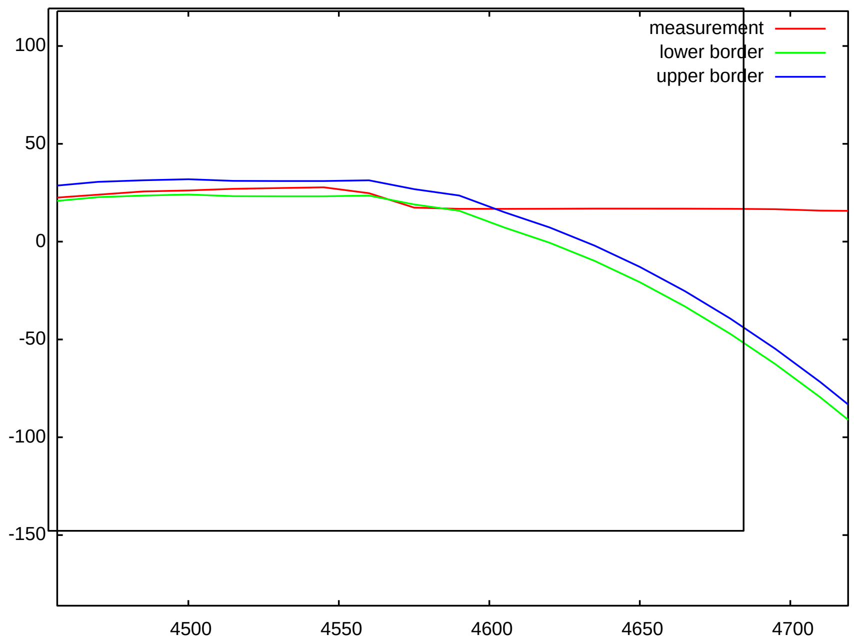

Figure 6: Instability of the algorithm when detecting a false negative.

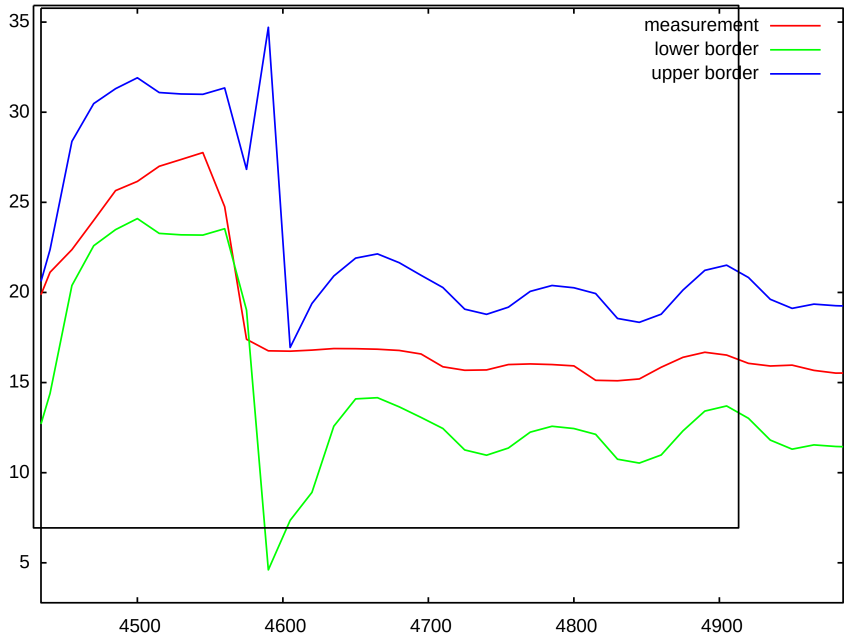

Figure 7: Instability workaround.

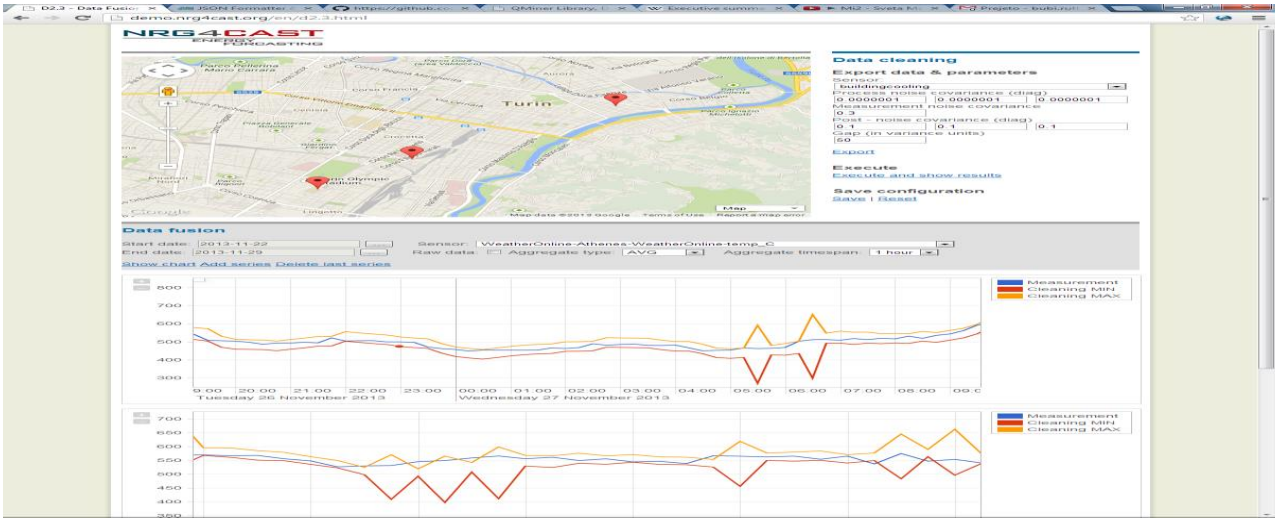

Figure 8: Illustration of the principle.

4.1.3.3 Proposed improvements

Proposed method has been very effective for detecting outliers, but there are quite some issues with the implemented approach, namely:

- frequently, there is a long-term error in the data (typically sensors are sending 0 measurement for a longer period of time),
- there are a lot of configuration issues with the Kalman filter,
- there is no mechanism in this approach that would detect missing measurements.

Solution should therefore be combined with a simple rule-based (or semantic) system that would be able to check certain sensor measurement against the sensible interval. Missing values can be handled later in the pipeline (with the merger components they can be interpolated or – with the complex event detection system they can be detected) – this is something that has been solved in NRG4CAST pipeline.

Difficulties with assessing the proper parameters of a Kalman filter can be diminished with the use of another algorithm. Simple moving average with a trend was almost as good as Kalman filter and it only needs one parameter (how much historic values will be used) to work.

4.2 Analytics

4.2.1 Event/Alert Processing Component

4.2.1.1 Concept

Complex event processing is a well known research area, where many possible solutions already exist. NRG4CAST project implemented an open-source solution on top of data layer and part of the application layer (mainly predictions) with the final goal to do some additional reasoning on top of this entire infrastructure. Additionally, CEP systems are able to perform some additional time-related operations on a stream, which can be useful in the simplification of the tasks within the pipeline. Another valuable experience has been gained with the alert distribution bus, see deliverable NRG4CAST D4.1 [9].

4.2.1.2 Evaluating Performance of Prediction Models

Two types of event patterns have been developed to deal with real data and prediction models. The first type of patterns can detect performance decrease of the model over time and the second type can detect spikes, when real live data are not captured in the prediction model. We solved the first problem with comparison of different performance metrics of the model over time. Spikes missed or wrongly predicted have been detected with the usage of autoregressive methods on differences. All these patterns could be easily expressed in the EPL language, supported by the chosen CEP system.

4.2.1.3 Lightweight Messaging System

When choosing messaging system for alerts we opted for a lightweight solution, integrated into the monitoring DB of the system. We used Websockets protocol to deliver the messages and a simple client/server infrastructure. The approach worked perfectly for fast prototyping, however we experienced some problems with the usage of this protocol behind the firewalls (like at the CSI premises).

4.2.2 Prediction Manager Component

4.2.2.1 Concept

Prediction Manager Component in the NRG4CAST setting is essentially a stream-mining solution. The main challenge in such a setting is fusion of multi-modal data from various sources and various types, and preparing it for use in stream mining techniques. Solutions have been described in detail in NRG4CAST D3.1 [10]. Various other innovations have been implemented.

4.2.2.2 Fusion of multi-modal data

When dealing with streaming setting one can encounter numerous small and obvious, yet non-trivial problems. How to deal with multiple data streams that are not aligned? How to deal with delayed data? How to deal with the data that is arriving in chunks, how to deal with forecasts (weather for example) or some static properties, that those models need for the time they are predicting a certain property?

Architecture and protocols to deal with such multi-modal data and how to bring it from its source to the final feature vector, used for modelling, have been developed within the NRG4CAST project and are part of

Prediction Manager component. Three basic types of data have been identified: sensor data (arriving in more or less classical streaming manner), weather forecasts – predictions (updated regularly), static properties (precalculated, but needed for a future time point). Modelling platform has been divided into two conceptual instances – the data instance and the modelling instance (see Figure 9). Data instance takes care about basic data fusion (calculation of aggregates and alignment) whereas modelling instance takes care about more complex fusion. An architecture has been proposed and implemented as depicted in the figure below. Many subcomponents that solve a particular subset of problems have been implemented: load manager, pusher, data receiver, local merger, resampler and meta-merger.

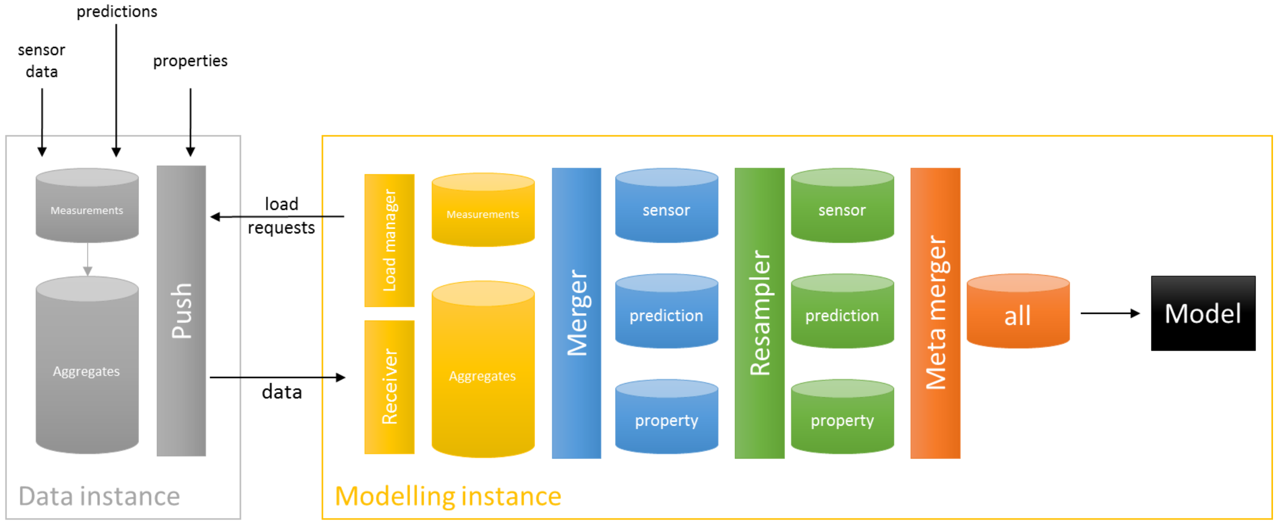

Figure 9: Data flow for modelling in the streaming data scenario.

4.2.2.3 Complex visualizations

Dealing with sensor data can be tricky. During development/initialization phase one needs total control over the available data, its derivatives and also ability for the introspection. Many visualizing components, linked directly to the data instance of the modelling platform, have been developed. They include: data availability visualization (time intervals of available data in the data instance), custom data comparison (compare different time series on the same/different scale and on the same/different chart), and exploratory analysis (visualize correlation between different pairs of time series).

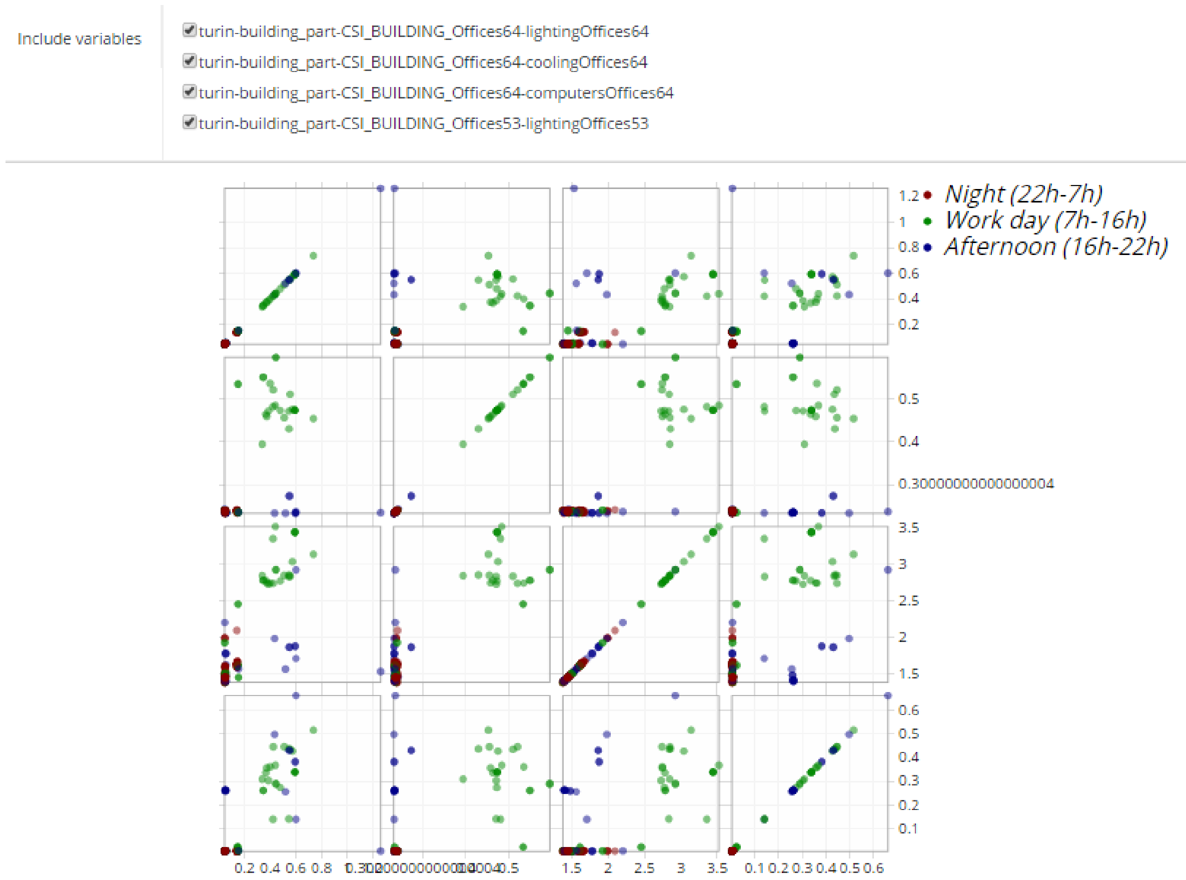

Figure 10: A drawn 4x4 scatter matrix, data points colored by hours in the day.

4.2.2.4 Establishing connection between evaluation and production phase

Selection of the models was traditionally done completely offline. With the development of stream mining methods the need of testing changed – as the methods expect the stream and not a chunk of data. The solution, depicted in Figure 9, also solves the problem of seamless transition between the testing and the production phase.

Testing phase uses the same mechanism for obtaining feature vectors as the production phase. The only difference is, that it is performed on a stream that is being pushed on-demand. This adds a nice but important feature to the system – the ability to use practically the same infrastructure for testing and production.

4.2.3 Energy Market Factoid Extraction

4.2.3.1 Concept

The initial idea, that was implemented and tested regarding Energy Market Factoid Extraction within the NRG4CAST project, has been proposed within NRG4CAST D3.2 [11]. It was based on a statistical approach within a well-known, but not very much exploited dataset of Google N-grams, the most common n-grams, occurring in the web. The n-grams were to be connected with energy market companies from Yahoo Finance and the most common quotes should be extracted and out of them patterns should be created. As depicted in Figure 11, these patterns should be aligned with Cyc Knowledge Base and the extraction of factoids should be performed on a stream of news, being additionally enriched.

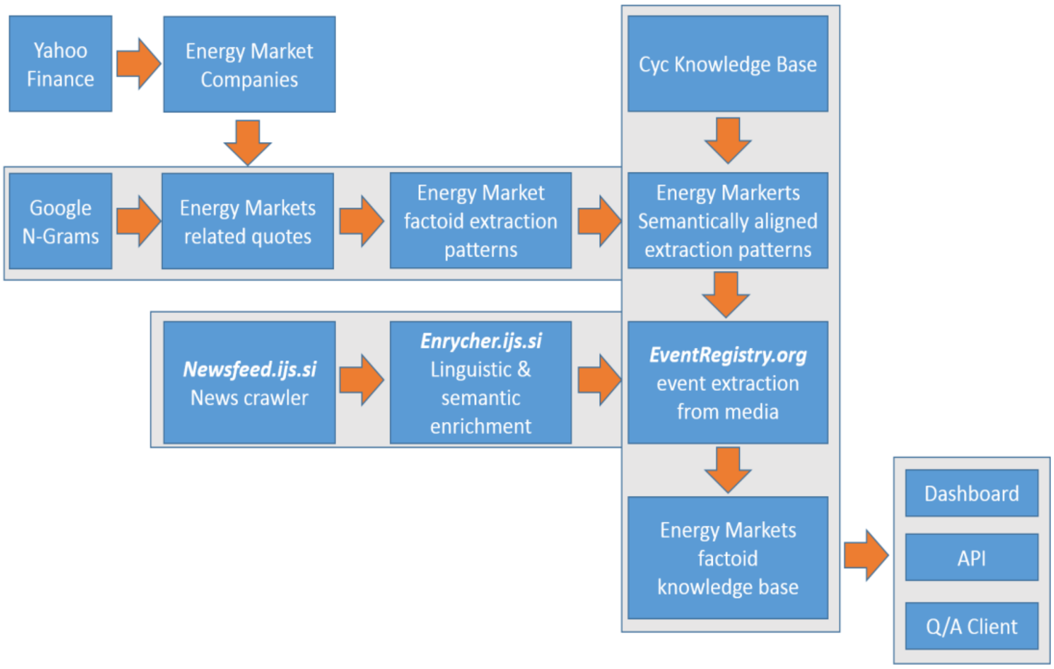

Figure 11: Initial concept for fact extraction within NRG4CAST.

4.2.3.2 Fast n-gram wildcard queries

The first innovation has been in creating a system that would enable fast wildcard search within the n-gram database. The solution is described in detail in NRG4CAST D3.3 [12]. Problem itself borders the big-data area. A special optimization with the usage for tries (prefix trees) has been developed, which enables the user to perform 10^5 queries/s using only one core.

4.2.3.3 NLP approach to the problem

The resulting patterns have, however, not been according to our expectations. Google N-grams are just not rich enough to be sufficient for the task at hand (many times even the name of the company includes several words). Statistical approach had to be changed with a standard stack of NLP tools including named entity recognition and dependency tree parsing. We proposed a methodology, which could actively learn from a stream of news and would be able to detect correlations within longer and more complex parts of the text.

Extraction of the patterns is illustrated in Figure 12. The outline of the idea is: we want to mark potentially interesting named entities in the text (Exxon and Mobil in the figure) and try to find the shortest possible path between the two in the dependency tree. The path would include relations as well as vertices. These relations could be transformed into patterns with a number for automatic or semi-automatic methods.

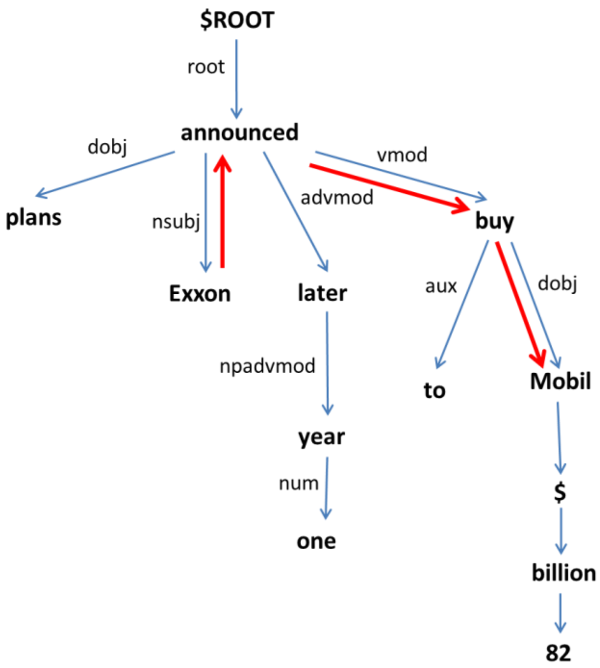

Figure 12: Example of a dependency tree with a possible pattern.

4.3 Visualization (CSI)

4.3.1 Concept

NRG4CAST visualization part consists of Real-time Visualization and Report Management & Visualization Components. Its responsible for the visualisation of the historical and real-time data coming from different sensors and visualisation of real time information related to NRG4CAST consumption centre, nodes and sensors, metadata and data related to external features. Moreover, the real time alerting is available within the Real-Time Visualization GUI [5]. The visualization part is responsible for the creation, management and visualisation of predefined and on the fly reports necessary for the NRG4CAST. The user can generate reports on different information and events from the sensor network logs, tasks, notifications, energy consumption and power usage.

NRG4CAST toolkit allows the user to visualize multiple data series and compare them to each other, navigate and select hierarchy of the consumption centres/nodes, visualize and compare aggregates for certain consumption centres or nodes for different time windows, create and export on the fly reports, visualize prediction vs. current and historical measurements, visualize prediction for different time windows (see Figure 13). The user can visualise a set of predefined reports on energy usage, consumption, savings and prediction. It also provides the possibility to monitor pilot overall energy consumption and external features such as weather parameters, etc.

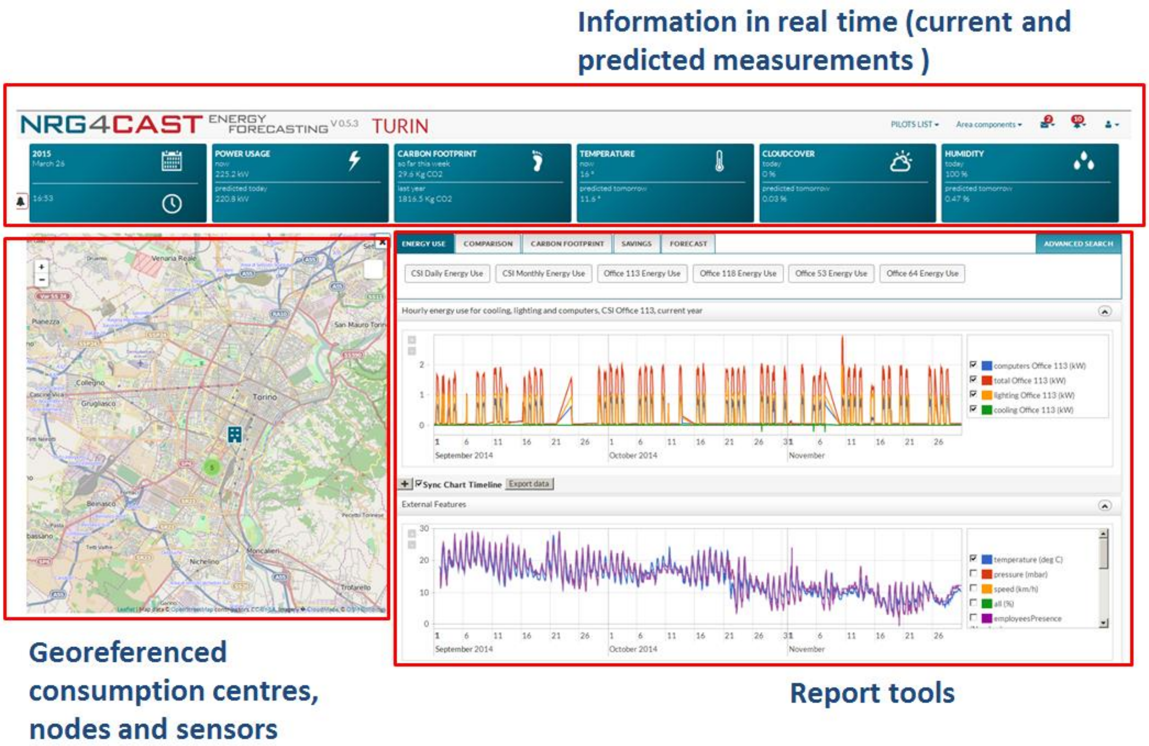

Figure 13: NRG4CAST prototype GUI.

It also offers the possibility to visualize georeferenced objects on the map, and to effectuate zoom on the area of interest (chosen pilot site). By clicking on the object, the user visualises all the information related to the consumption centre, node and sensors installed (metadata, location, selected sensors last measurements). The alarm list is connected to the geo-referenced objects. An alarm icon appears on the object once an alerting notice is registered. The NRG4CAST visualisation part is described in detail in Deliverable 6.4 [5].

4.3.2 Problems faced and solutions found

4.3.2.1 System performance

The long-term time series predefined reports were requested by pilot partners. These long periods reports include considerable amount of data which requires time-consuming data processing. The response time in this case isn't compatible with the online application performance.

Many of these huge reports are static. As an example, visualization of offices electrical energy consumption within the last year. In these particular cases the cache solution was applied at the system visualization level. This solution has significantly improved the system performance and reduced response time.

4.3.2.2 Dynamic configuration

During the initial analysis of the use cases we noticed many similarities between the various reports required by the pilots. Most of the reports and functionalities of NRG4CAST pilot prototypes GUI have the same logic. Often only a few parameters are subject to change.

During the implementation phase, the services, data structure and hierarchy were continuously changing. So, we came up with a solution that would allow for dynamic configuration of the application using a configuration file json. The Report Management & Visualization and Real-time visualisation components are characterized by dynamic configuration. This choice gives a possibility to easily add and remove reports and functionalities within the prototype.

This flexible solution has allowed for implementation to a number of pilots and numerous reports starting from a single application and to manage the layer of visualization and reporting in a more effective way.

Pilot configuration includes:

- pilot description,
- databases,
- prediction services, prediction model type,
- alerting services,
- other technical parameters (cache, time out, …).

Configuration of reports *includes :*

- sensors,
- weather stations,
- weather parameters,
- report type,
- simple reports (a number of phenomena and various time spans),
- complex reports (savings, forecasting, carbon footprint).

4.3.2.3 Complex configuration of pilots

One of the main problems we faced was the growing complexity of reports. Many of the operations and functionalities were unforeseen by the initial pilots version and appeared during the implementation stage. We could overcome this obstacle by redesigning a part of the configuration engine.

Actual configuration is stored in the static JSON file. The future solution could be an interface configuration managed through the monitoring database. Furthermore, implementation of a back-office for functions management would give a possibility to create a GUI able to modify, add and remove reports, add functions, etc. This solution increases the system complexity but also would make NRG4CAST system more solid and flexible.

4.3.2.4 Websocket

NRG4CAST project uses the WebSocket communication Protocol10 for monitoring and alerting data in real time. At present, some proxy's and firewalls don't fully support WebSocket specifications. WebSocket foresees two-way communication, while proxy and firewalls are historically meant for traditional client server communication. In order to overcome this issue, the two solutions were implemented within the NRG4CAST viewer for alerts retrieval. The first solution uses WebSocket protocol. The javascript client contacts the server websocket ngrEventBus and receives alerts in real time. Another solution uses http calls sent automatically by client to the viewer server. Actually, in the latter solution, a WebSocket client is installed as a component in the Data Access & Integration platform and continuously stores alarms in the Monitoring database. The viewer server then queries the monitoring database every 30 seconds through the Generic SQL query service of the Data Access & Integration platform in order to retrieve the last alarms.

4.3.2.5 Real time data visualization

The visualisation of information in real time and the possibility to highlight certain most important measurements was requested by the project partners. Moreover, it was necessary to find a solution for visualisation of alerts in real time. A unique, immediate and straight forward user interface adapted for various user profiles was also requested.

Based on these requirements we designed the NRG4CAST toolkit as a dashboard (see Figure 13). The upper panels are referred to the whole pilot and represent current and predicted measurement such as:

- current and predicted energy consumption,
- carbon footprint so far for the current week and predicted for the next year,
- temperature now and predicted for tomorrow,
- cloud cover now and predicted for tomorrow,
- relative humidity now and predicted for tomorrow.

Furthermore, the alerts in real time are visualized on the map. Through this way it is simple and immediate to associate consumption center /node and sensor with the corresponding alert. This solution addresses also to users with less technical profiles.

4.3.2.6 Icons to simplify NRG4CAST GUI

While implementing the NRG4CAST toolkit GUI, we've realised that a lot of information was to be visualised at a single page. To improve the dashboard usability we used icon that would simplify viewing and give straight to the point information avoiding the use of long space consuming labels. We tried to assure the consistent use of icons. Normally one icon, designed in a certain way, has the same meaning within any part of the prototype.

Additionally, we relied on the use of colours which has proved to be particularly useful to highlight the various alert levels.

4.3.2.7 Advanced search

During the project implementation, we realized that often we do not have a clear picture where the meters are installed and what exactly it is measured. So, we developed an advanced search option, which gives the user the possibility to easily navigate within the sensors hierarchy and to locate the sensors of interest (see Figure 14). There are options to select consumption centres, nodes and sensors to be visualised on the graph for a certain time frame and also analysed. All chosen data can be exported in tabular format.

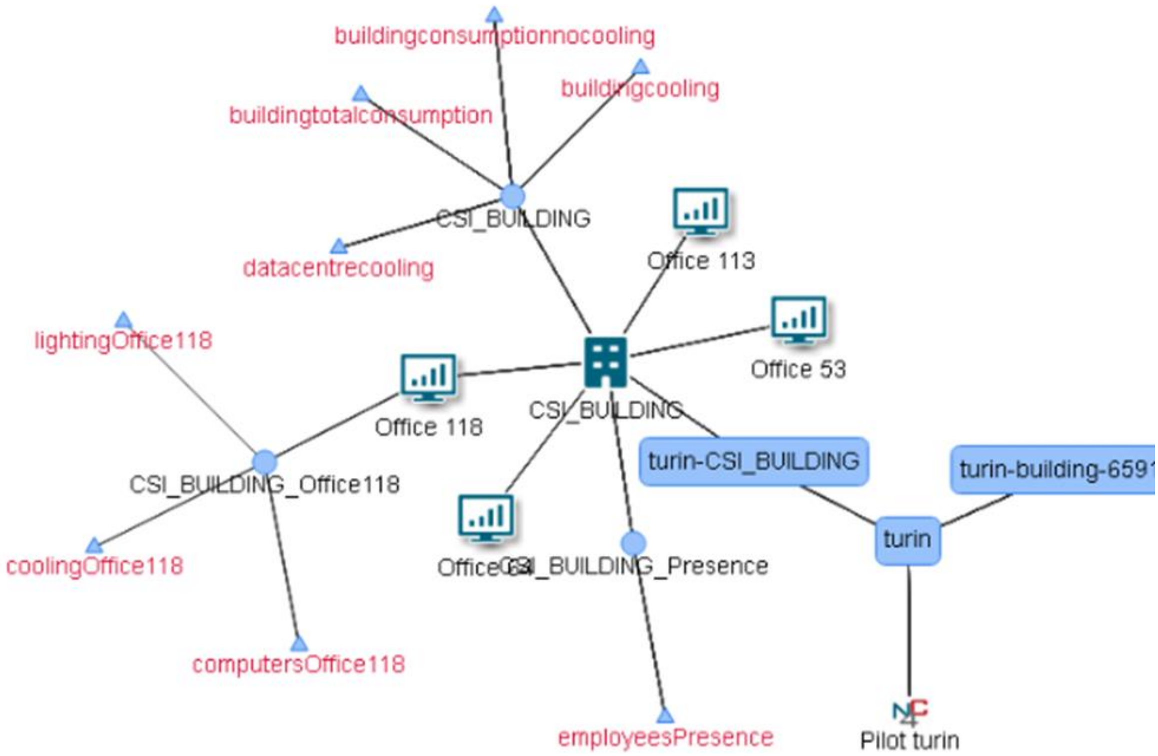

Figure 14: NRG4CAST hierarchy visualisation example, Pilot Turin.

4.3.2.8 Set of predefined reports

One of the main requirements for NRG4CAST GUI was a visual representation of energy use and energy consumption (historical data, real time data and forecast). A set of predefined reports was requested by each pilot. The predefined reports are an effective way to change consumption patterns and to compare energy use with a number of parameters.

In order to meet these requirements we developed a set of predefined reports on energy consumption, possible savings related to various actions in kW/h and forecast. These reports provide visual representation of energy consumption trend within a certain timeframe. Once report diagram appears at the screen, there is a possibility to compare energy consumption, energy usage and thermal energy production with a number of parameters such as temperature, pressure, humidity, wind speed, wind direction, cloud cover, precipitation, visibility, people presence at the building etc (refer to Figure 15). Moreover the user can visualize multiple data series (by unchecking Sync Graph Timeline) and compare them to each other. The propuced reports are exported in tabular format.

A tool for visualisation of a set of predefined reports is a central part of the NRG4CAST GUI. We tried to gather all the pilots requirements and to satisfy their needs. The developed solution is effective and quite complex and gives the possibility to the technical administrator to add various types of reports, to define and modify the reports content.

Figure 15: A set of predefined reports.

4.3.2.9 User cases for pilots

At the non-technical level we faced some communication problems. We noticed that communication only by email sometimes was not sufficient. So, it was decided to organise quite frequent skype and gotomeeting b2b meetings with partners in order to discuss the pilot use cases and to define the pilot layout in detail.

4.3.2.10 Webservices

A number of webservices was implemented to provide data to the NRG4CAST system. The REST services were implemented to provide aggregates of measurements for one sensor or a number of sensors to be used by NRG4CAST visualisation components. There are API for storing, updating and retrieving predictions etc. These services were modified during the development of the system and we faced some difficulties caused by lack of planning and scheduling in services publication and editing.

We overcome this problem by creating a project living document NRG4CAST configuration log, continuously updated.

4.4 Authentication and authorization

4.4.1 User & Role Management Component

4.4.1.1 Concept

Authentication is the act of confirming the truth of the identity of an entity, usually of the user of a system or of a specific data item. Authorization is the function of specifying access rights to resources for a specific entity, i.e. to control access of the users to the resources and the functionalities of an information system. A common scenario for access control is the aggregation of users into specific groups and roles and controlling access to system resources and functionalities according to the assignment of permissions to the role pertaining to a user.

4.4.1.2 Problems faced and solutions found

The main issue related with Authentication and Authorization of the NRG4CAST toolkit users stems from the fact that the NRG4CAST toolkit consists of a number of loosely coupled components running on different server processes. Specifically, the user should access the multiple components by providing his/her credentials only once. This process in known as Single Sign On (SSO)**.** There are several open source solutions for adding SSO functionality to software components. After a comparison, the Central Authentication Service (CAS) has been selected as it is provided with an Apache license which is permissive for the commercialization of the NRG4CAST toolkit, it supports a plethora of different software technologies (Java, .Net, PHP, Perl, Apache, Python, Ruby etc) and it is widely used and well supported.

The flow for authentication is depicted in Figure 16 and can be described by the following steps:

1. The client attempts to access the NRG4CAST Web UI of a component
2. The application redirects it to CAS. CAS validates the client's authenticity, usually by checking a username and password against the User/Role Database.
3. If the authentication succeeds, CAS returns the client to the component, passing along a security ticket.
4. The component then validates the ticket by contacting CAS over a secure connection and providing its own service identifier and the ticket. CAS then gives the application trusted information about whether a particular user has successfully authenticated.

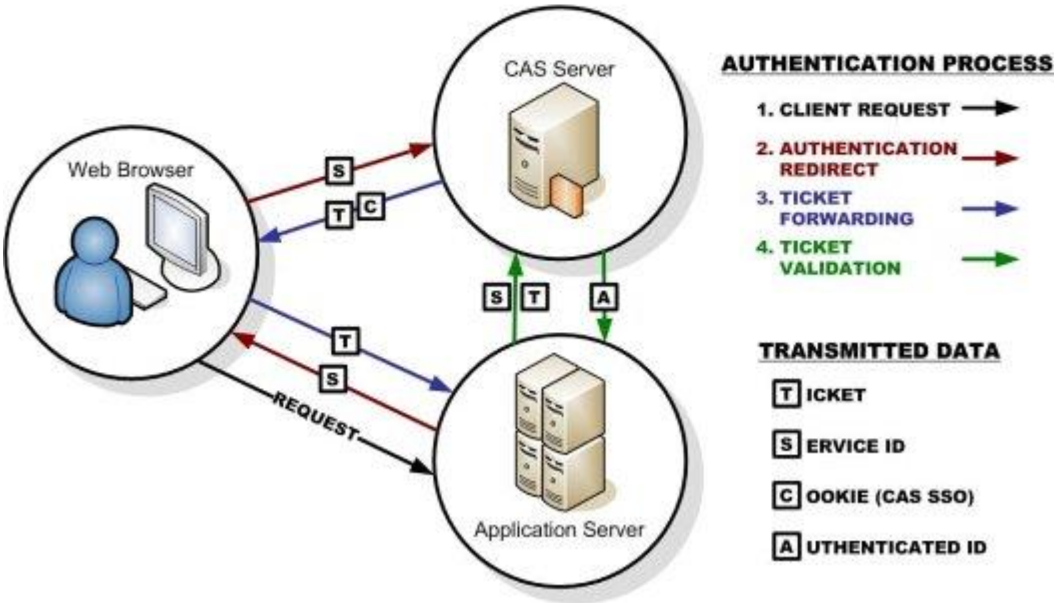

Figure 16: CAS Authentication flow. [13]

From the above description it is evident that CAS is responsible only for authentication and not authorization of a user to the functionality of a specific component. Each component handles authorization internally: the security ticket is retrieved concerning the authenticated username. Then the relevant component retrieves the role of the user from the User/Role DB and provides or rejects access to relevant functionalities autonomously. Therefore, authorization functionality should be implemented independently on each different component due to the loosely coupled nature of the NRG4CAST toolkit.

The User/Role DB is designed according to the NIST standard for Role Based Access Control [14].

5 Pilot cases implementation, lessons learnt, recommendations

The aim of this Chapter is to summarize the lessons learned and recommendations from the usage of the NRG4CAST toolkit for pilot cases. The recommendations presented in the following table are based on the experiences gained from the realisation of 5 different pilots in the NRG4CAST project from December 2012 to May 2015. Those pilot cases are, namely: the Smart charging algorithm pilot in Aachen, Germany; the Light as a service pilot in Miren, Slovenia; the energy savings pilot at the University Campus in Athens, Greece; the energy efficiency pilot in public owned buildings, mainly public offices in Turin, and the 2 pilots in Reggio Emilia, Italy: energy efficiency of the production of district heating and energy efficiency in the case of thermal energy consumption. In the Table, the various challenges and problems the pilot cases were faced with the solutions found and lessons learned are stated. It can be seen that the challenges and problems can have a completely different nature and severity from one pilot case to the other.

Table 1: Lessons Learned from NRG4CAST pilots implementation.

Challenges and problems to be solved	Lessons learned and recommendations
Energy savings at the University Campus in Athens, Greece	
Initial measurements from the sensors installed in Lampadario and Hydraulics buildings were uploaded to the FTP server of the NRG4CAST toolkit on a daily basis. The upload process was initially manual due to the limitations of the NTUA sensor equipment software. This limitation was overcome in the case of the new sensors which were installed in the context of the NRG4CAST project.	The establishment of an online communication link between the data captured in NTUA and the partner SingularLogic proved to be a challenging task. The problems related not only to the technical part of the communication, but also to the confidentiality of data. Fortunately, all challenges were adequately overcome and the data have been communicated and used for the NRG4CAST purposes. An ftp has been created through which the data from the whole campus are sent online to SingularLogic.
During the final roll-out of the sensors installation the following were installed in the electricity substations of the campus so as to cover the whole campus (62 buildings): - 34 electricity meters SIEMENS SENTRON PAC3200, PAC3100 - Schneider controllers 701,711 and Schneider module XENTA 451A - 16 electricity meters Schneider Electric PM3250 - 4 temperature sensors & relative humidity sensors (KIMO TH210-BNDI-150) - 4 lux meters Vernier LS-BTA	During the installation process, it was realised that the grid architecture of the campus substations was outdated and that the same lines were be taken to whole buildings or even to block of buildings. Only the electricity lines of the Mechanical Engineering Department buildings could be split to individual loads, like lighting, HVAC systems, and other operational loads. Thus, the use of Schneider controller 711 will allow multifunctional system management control with access to measurement and control of these applications (lighting and HVAC). Following the performance of suitable programming of this installed controller, lighting and HVAC systems as well as other operational loads being identified will be controlled.

Challenges and problems to be solved	Lessons learned and recommendations
	Consequently, apart from real time data, energy management opportunities are identified for the campus.
The 2012 Energy Efficiency Directive establishes a set of binding measures to help the EU reach its 20% energy efficiency target by 2020. Under the Directive, all EU countries are required to use energy more efficiently at all stages of the energy chain from its production to its final consumption. Measures have to ensure major energy savings for consumers and industry alike.	Energy Saving Senate Committee of NTUA was responsible for the submission of several national proposals for energy saving measures applied to NTUA campus buildings. The Energy Efficiency Certificates of Campus buildings were necessary to be conducted for the submission of these proposals. NRG4Cast platform may replace the energy codes based on Energy Performance Directive used for the EPCs and reduce the time and costs of Energy Performance Certficates.
Energy Service Companies (ESCOs) have been operating within the energy sector for many years and recognise the challenges that block of buildings face and the need for change. With the technology, expertise and proven solutions to help develop a strategy outlining the most effective path to improvement, they take a comprehensive view of a building's carbon footprint, identify the areas where improvements can be made and implement practical, engineered solutions to reduce energy consumption and carbon emissions. The benefits are immediate, including reduced costs, improved building comfort and legislative compliance.	NRG4Cast is an important tool for ESCOs. Through an Energy Performance Contract NTUA could: • Reduce the financial risks associated with energy consumption • Utilise ESCO design, implementation and finance resources to improve the energy efficiency of buildings • Conduct a detailed energy audit to identify where and how much energy demand can be reduced Reap guaranteed cost savings. Energy savings are guaranteed by the ESCO. In the unlikely event of the agreed savings not being delivered, the ESCO makes up the difference.
Demand response is an electricity tariff or program established to motivate changes in electric use by end-use customers, designed to induce lower electricity use typically at times of high market prices or when grid reliability is jeopardized. In regions with centrally organized wholesale electricity markets, demand response can help stabilize volatile electricity prices and help mitigate generator market power. Demand response can include consumer actions that can change any part of the load profile of a utility or region.	NRG4Cast by using different tariffs could affect the demand response of NTUA campus by changing the load profile leading to less energy costs.

Challenges and problems to be solved	Lessons learned and recommendations
Light as a service, Miren, Slovenia	
Visualisation and monitoring need to be simple and understandable for different type of user.	On the first screen (page) of the tool put maximum 5 basic (very significant) information.
The performance monitoring	The performance monitoring with different KPI must be included.
Technology from bottom to top	Technology from bottom to top needs to be specified in reason that the user could understand what they need to deploy to get the result.
Definition of Business models	Business models need to be very fine detailed to understand the benefits.
"From concept to realisation "	The implementation period from concept to realisation needs to be defined through business model.
Smart charging algorithm, Aachen, Germany	
Use a top down approach	First of all, the top requirements of the stakeholders must be identified to establish further actions. It might be good to have ideas about what can be done with the data (bottom up approach), however it is crucial to prioritize those ideas in close collaboration with the stakeholders. From that prioritized list, the consecutive steps can be derived. For example, the grid provider has in short notice a higher priority for monitoring the return of investment of their charging stations. Therefore, monitoring the overcharge of the charging station network is only ranked second.
Do not underestimate the data transfer	Since the original approach planed to deliver data from a certain amount of vehicles, one crucial part was to establish a working data transfer via a mobile network. The project showed, that the technical problems with such a stable data connection are more complex than estimated. Especially the amount of data (large data bits that are send quite often, add up to a large amount) was too large for the network to handle. In addition, the network coverage was not as good as expected. A

Challenges and problems to be solved	Lessons learned and recommendations
	wired data transfer should still be preferred.
Test series must be large enough and accessible	It is essential to monitor large amount of data to make precise forecasts. Therefore, the number of vehicles was essential. However, a large number of vehicles would imply a huge investment for gaining access to vehicles and modify them accordingly.
Test series need to be in your control	Besides having access to a large test group, it is essential that the test group must be under our control for a successful validation. In the case of electric vehicles, it was not possible since only one car could be controlled directly. Other vehicles, that are monitored could not be accessed directly, because the vehicle was officially owned by a partner. Therefore, the monitored vehicles are used rarely. As a result only small amounts of data had been delivered.
Data cleaning and validation is essential – “do not trust the sensors”	Quite often, sensor delivered invalid or missing data. Consequently, there should be several validation series executed to clean the data. For example, it can be monitored by a car, how much it charged and afterwards compare the value with the value of the charging station. Additionally duration and location can be easily validated. Furthermore, rules can be defined to separate wrong data values from realistic ones. For example, a value interval can be created to restrict the data stream to realistic values.
Energy efficiency, production of district heating, Reggio Emilia, Italy	
The pilot scheduling did not take into account the period of thermal energy supply, that usually starts in mid October and ends in mid April. This means that the platform or at least the first prototype should be ready before the thermal energy supply periods begins.	It is important to plan pilot case studies according to the specific features of the partners that implement it and to the availability of data.
The role of IREN is to provide energy data, but IREN is a service provider, and often the energy manager that handles the service to the final customer is a different one. This means that the collection of energy data of real users could be difficult or not possible.	It will be useful to evaluate the availability of data directly at the stage of proposal writing, in order to set up a pilot where the needed data are directly supplied by the provider involved into the project.

Challenges and problems to be solved	Lessons learned and recommendations
In very wide networks or service infrastructure it is very common that the hardware is different in many parts or sub-parts of the network. This happens where there is no strict commitment with a hardware supplier. Hence sensors, costers and other interfaces could be totally different. As a consequence also data formats are different. This means that extra work is needed to gather and clean the data.	When the partner knows that the data it has to provide to the project have different formats, it will be useful to plan some effort for data cleaning and reformatting directly at the stage of the proposal preparation. Otherwise it is important to carefully set up the pilot in order to encompass that part of the network that has the same hardware.
Sometimes data to be provided to the tool are not sent in a consistent time basis. For example the sw coster allows to set the data recording with a range from 5 minutes to 240 minutes. But memory resources are limited: this means that if the record is done once an hour, every week data should be downloaded by the operator, otherwise they will be erased. If the record is set every 10 minutes, data should be downloaded every day. If the operator work time is off, data are lost.	It is important to define policies for risk mitigation, in order to get data independently from operators availability at work.
In this project IREN is a technical partner, providing the infrastructure on which to test the case study (namely the pilot). As often occurs with technical partner, IREN has a lot of technically- specialized employees and less project managers that cope with EU-funded projects. This situation can lead towards ineffective management of effort and planning.	It is important to involve from the beginning of the project managers with a long-lasting experience in the management of EU-funded projects.
IREN is a big company that encompasses several sub-companies. Sometimes IREN employees shift sub-company, even if they still will keep the same role/professional profile. This situation creates some problems in managing activities and financial reports.	It will be useful at the stage of proposal preparation to include different sub-companies of a big company or holding.
Energy efficiency in public owned buildings, mainly public offices, Turin, Italy	
Validation of project results. Within the first 12 months of the project it was not possible to prove the project results, at least 24 months were needed to show achieved results.	According to Directive 2012/27/EU transposed into Italian Decree 102/2014 in July 2014, starting from 2014 and until 2020, public bodies will have to progressively reduce the energy consumed on their own premises. ICT applications can significately reduce energy use in buildings and lower greenhouse gas emissions. In the case of existing public offices in Turin, NRG4CAST system acts as a solution for energy efficient and sustainable ways of managing and

Challenges and problems to be solved	Lessons learned and recommendations
	using public owned buildings. Starting from the beginning of the project it is crucial to calculate the projects indicators in order to be able in the future to analyze significant deviations (such as project profitability, Economic savings, Energy savings, CO_2 savings).
System requirements: It is problematic to find a unique solution for all the pilots.	We tried to respect the individual pilots requirements during the system design by gathering requirements for each pilot. Through this way NRG4CAST respects the individual needs and gets higher acceptability of measures. The project Scope Matrix, the living document with the project requirements is updated on a regular basis.
Show an early draft of the prototype to the stakeholders.	It could be wise to show a preliminary system prototype to the stakeholders. It is always good to have "an eye" from outside of the project. This gives an opportunity to have a first feedback and to make important changes. Stakeholders/future users are to be involved from the very beginning of the project. User active participation is crutial for the system implementation. It is important to record user feedback and suggestions, negative and positive perception. The participatory workshops proved to be a very effective strategy for sharing the project early results. These workshops help to develop and raise awareness of the stakeholder while on the other hand it gives stakeholder the possibility for involvement and engagement in the NRG4CAST project.
ICT for energy efficiency at the public offices in Italy does not function without employees' active participation.	Only motivated staff can improve functionalities of installed systems. Motivation system has to be a part of the project design in order to involve employees. We tried to raise awareness among employees using website and intranet facilities, posters and totems. The Awareness Campaign for Turin pilot consisted of the following actions: ✓ Involvement of employees, additional focus on personal benefits ✓ informing employees on scientific / technical issues of energy saving and

Challenges and problems to be solved	Lessons learned and recommendations
	energy efficiency ✓ encouragement towards concrete action ✓ visibility on the obtained results ✓ facilitate sharing of Best practices also outside of your building ✓ maintain Best practices and habits in time ✓ gather feedback on the initiative and suggestions for future initiatives
Sensors naming, no standard was applied.	NRG4CAST system integrates data from 5 pilots. Each pilot provides a vast set of sensors to the project. It would very useful for the system development, mostly for the visualization part, to set the standard for the sensor names. It would give a possibility to make the visualization part more flexible and easy to configure.

6 Further research areas

NRG4CAST brought together many different techniques, known in different areas of the data mining/ machine learning community and also among the Internet of Things community. Main research areas worth investing would be focused around: data cleaning (improved Kalman filtering), knowledge generation (on top of models), transition of the models and process modelling – namely for visualizations.

Kalman filtering improvement is needed because as we found out it is rather difficult to fine-tune many of the Kalman filter parameters to make the algorithm useful for different kinds of sensors. Research that might reduce the number of free parameters (e.g. estimation of the original hidden state is not needed) is already available [15]. The Impact of such an algorithm on the implementation and usability should be studied.

The most fruitful research area would be the area of **reasoning and knowledge generation** on top of existing streaming models (either for prediction or for root cause analysis). The most common goal would be to develop a methodology for optimal model selection, based on selected markers over the observed time-series. A lot of such research on meta-classifiers or on the usage of ensembles of prediction methods exists already (for example [16]). To use such a methodology of course the platform should be able to run multiple models simultaneously. A lot of reasoning can be done on top of the prediction models and actual data to support the Decision Support System. NRG4CAST has implemented some: detection of model deterioration and detection of sudden unhandled peaks of the models. A system could be simply expanded with a proactive event detection (based on existing rules, but evaluated on predicted data).

A lot of research on knowledge generation should be focused also on the area that was pioneered within the NRG4CAST projects (with collaboration with ProaSense project) and this area is Root Cause Analysis. This novel approach was introduced at the Discovery Science conference [17] in 2014. As the results are still experimental and as it seems that this might be a wider research area, there are a lot of opportunities here. One might study the impact of different existing clustering algorithms, a lot of research is needed in the area of feature generation and there is also a lot of opportunities in developing a customized clustering algorithm like the one that was reported in NRG4CAST deliverable D4.3. As such a system is quite complex a lot needs to be done in the area of visualizing and discovering useful results.

The last area connected with modelling is the area of generalizing the models. It is a research area that would need a lot of data for testing, but the outcomes could be really useful. For example – with IREN proposition (see below) one would like to build a mathematical (process based) model of energy consumption for buildings. The current state of the model is rather simple, based on 2 measurable and 1 customizable parameter. Model could be expanded with relevant parameters (see below) and could use any of the known prediction methods. Such a model would be a general one and could be transferred to different buildings, places, areas where no sensor data is available otherwise.

In the area of **textual pillar** of the NRG4CAST platform the following interesting research topics have been discovered: acquisition of fact extraction templates for the news data and nowcasting. Acquisition of fact extraction templates has been tackled from many different directions in the NRG4CAST project, but with a limited success. The research area behind this problem is a vast one and quite developed. It requires a separate big-scale research in a form of a project not only as a part of a project. The idea of nowcasting is not a new one. The term means that from the current data we want to estimate certain well known measures (for example GDP) which will be calculated in the future for the "now". These parameters could be already estimated on a basis of a measurable phenomenon (from the news, Twitter, etc.). Such parameters could greatly improve certain models (for example for energy price or for energy consumption on a spot market level).

The last area is linked to the more technical aspects of the platform. During the development of the platform we have found a need to be able to model processes, linked to sensors. This means, that we are able to transform the sensor value through a certain functions or to define a function composition or similar. Formally this is a very well researched and standardized area [18], however in practice not many solutions exist that are able to implement these features. For the reporting this is one of very basic functionalities that can greatly improve the stakeholder's experience with the system.

Another important topic nowadays is energy and in fact there are two major fields of interest related to **energy markets**: the first one regards to ESCO projects and performance contracting and the second is energy grids.

With respect to ESCOs and performance contracting there are two critical aspects to be solved. The first one regards future prediction of energy and financial success of the projects results. The second one is the verification of the project results (performance contracting) with respect to monitoring and verification utilizing the minimum number of sensors (reduction of M&V costs). For the first case it is of high importance to accurately analyse current project parameters yielding to specific energy consumptions and to estimate future energy consumption or energy savings (i.e. financial savings) with respect to certain technological interventions where their performance relies in a series of variables. A typical example of the above is the prediction of the buildings heating and cooling loads and performance of HVAC systems where, in return, they are influenced from system variables irrespective to building loads (such as aging of equipment, seasonal efficiencies, etc.). For the second case, namely M&V, the minimization of the number of sensors can be solved through the development of mathematical models which calculate energy consumption or energy savings with minimal monitored parameters (e.g. temperatures, time and electricity consumption).

With respect to energy grids and the emerging dynamic pricing markets, balancing supply and demand is of particular interest to be investigated. In order to achieve that, monitoring of energy consumption and information extraction of the energy production are of crucial importance. In order to take advantage of dynamic pricing markets and at the same time to sustain grids stability, demand and supply have to be accurately predicted. A major challenge for this particular case is to develop prediction models, particularly from the demand side, where loads are fluctuating in a dynamic manner and grids have to respond on this particular demand, often partially served by external unstable sources (such as on site installed RES), and all the above parameters, as economical and environmentally sound as possible.

Another possible future research area is the concept to a refined **model for energy prediction**. In the last years, research in forecasting models and performance improvement is very active: usually, these targets are achieved through the installation of extra sensors, devices and equipments able to provide the data needed by the computation model. Unfortunately this is a crucial limit for the exploitation of this achievements. In facts today's market trends force companies to cost reduction, that often means hardware reduction.

An alternative to the installation of new sensors is the deployment of Add-On Functionalities (AOF). They do not require new sensors but only information coming from the network and its sub-systems. The information is computed in a well tuned algorithm and the results provide an added-value in terms of cost reduction and production start.

In NRG4CAST project we decided to open a new research issue, investigating the possibility of predicting thermal energy consumption by calculation. We applied this prediction model to Campus Nubi pilot, in Reggio Emilia City. First results are encouraging, but a lot of work should still be done. What should be better investigated and refined is the variable affecting the energy rate coefficient. Since this research is very demanding, the experience gained with Campus Nubi pilot paves the way for a new big research project at European level.

In fact, the computation for the thermal energy prediction is really simple, as elicited in the following figure.

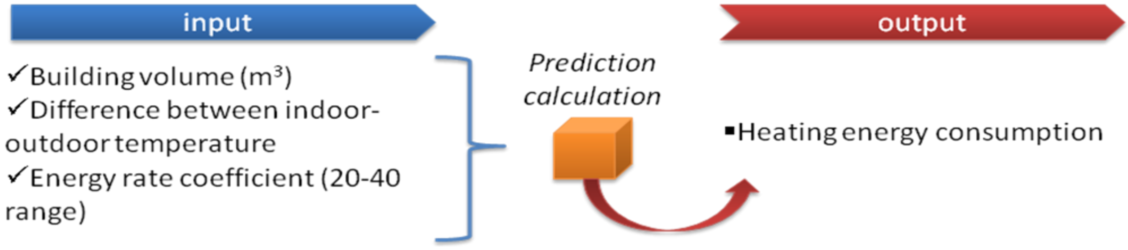

Figure 17: Input and output for thermal energy prediction.

Heating energy consumption (W) = Building volume (m^3) x indoor - outdoor temperature difference (C°) x Energy rate coefficient (range between 20 and 40)

What is very challenging is how to establish exactly the energy rate coefficient. Up to now the only way to establish it is through heuristics and personal experience of thermal engineers.

What we had the chance to investigate in NRG4CAST project are the factors (independent variables) affecting the energy rate coefficient. The idea behind is to have the premises to start another project focusing on a model of the energy rate coefficient. Each variable and the relationships and constraints between variables should be in depth studied, in order to achieve a theoretical model that, by means of a proper algorithm, allows to detect the precise energy rate coefficient for each building.

The aim of the prediction model for energy consumption was twofold:

Providing energy consumption estimation, allowing in this way to use the NRG4CAST forecasting tool without adding extra sensors to the current energy network.

Contributing to improve the efficiency of the thermal energy production taking into account the Adaptive Thermal Comfort Standard.

The “thermal comfort standard” is determined by different factors (i. e. internal air temperature, relative humidity, ventilation, etc.). For example, the suitable temperature varies according to many factors like the type of activities that are going to take place in the building.

The “thermal perception tolerance” is a key element to define the thermal comfort standard range of a room or a building. This means that the internal temperature value should be an adaptive parameter instead of a fixed one.

The set of independent variables to be investigated and modelled both for the energy rate coefficient and for thermal comfort standard is the knowledge base (KB) the prediction algorithm will use. Within NRG4CAST project the KB variables have been identified. However, the KB in-depth definition and implementation is out of the scope of this project. The important parameters that influence the 2 coefficients are listed below.

The Energy Rate Coefficient is influenced by two types of variables:

1. Physical environmental factors
    - ✓ Solar irradiation (W)
    - ✓ Wind speed (m/s)
2. Artificial design parameters
    - ✓ Transparency ratio (%)
    - ✓ Building form factor
    - ✓ Orientation
    - ✓ Properties of materials (walls, floors, roof, ceilings, windows, doors)

- ✓ Distance between adjacent buildings
- ✓ Infiltration load
- ✓ Ventilation flow

The Adaptive Thermal Comfort Standard is influenced by two types of variables:

1. Lifestyle
   - ✓ Lighting
   - ✓ Equipment & furniture
2. Thermal perception tolerance
   - ✓ Use of HVAC (Heating, Ventilating, and Air-Conditioning systems) vs NV (Naturally Ventilated)
   - ✓ People load
   - ✓ Degree of activity vs typical application (i.e. seated, standing, eating – restaurant; dancing – dance hall)

Each of those variables should be modeled and weighted. A user-friendly tool could be developed, considering both web services and mobility needs (i.e. realizing App for mobiles). The integration of the energy prediction model with NRG4CAST platform has also been evaluated so a future research area has been highlighted.

7 Conclusions

Quite a number of tools regarding energy monitoring, energy management, and future energy demand estimations exist in the market nowadays. The NRG4CAST concept was to develop a powerful tool able to deal with big amount of data, different data origin (web data, sensors data, etc), different data time series, and provide with a specific toolkit which integrates different solutions and information sources and a decision-making assistant for short-mid-long term planning of energy demand. The developed NRG4CAST tool provides with a variety of services which monitor energy consumption on different levels of interest, mechanisms that control energy demand, alert in case of anomalies or faults, accurate prediction features for energy demand and energy prices.

The NRG4CAST tool is a platform providing with complete energy services related to monitoring, alerting, reporting and forecasting. It is custom made so that the user gets exactly the information he needs. The tool has the ability to be extended to other important fields like water, environment, and industrial processes.

The tool would have not been completed without the collaboration, continuous assistance and the strong belief for the final outcome from all the participants. Especially the technical partners showed particular interconnection skills and contributed determinatively to the completion of the computational methods and the final shape of the NRG4CAST tool. However, the road to reach such a result was not simple or straightforward. Many obstacles had to be overcome. Regarding the latter, the purpose of this guide was to gather all the experience gained through the tool's preparation phases. In view to that, the Best Practices report is providing, in a compact way, the problems found, the lessons learnt, and the knowledge gained, along the duration of the NRG4CAST project, with regards to the main NRG4CAST components:

- Data collection
- Analytics
- Visualization

enriched with the Authentication and Authorization features.

The report provides with information about the obstacles/ problems that were encountered during the implementation of the different sub-components, the solutions found, and it presents the acquired knowledge. The experience gained from the implementation of the tool in real case studies has also been recorded since it provides useful information for the inter-linkage obstacles faced with during the operability of the tool in connection with the different data input sources.

Finally, some topics for further research areas are pointed out regarding machine learning issues, reasoning and knowledge generation, together with some thoughts to broaden the NRG4CAST features.

References

[1] "Data Integration", Wikipedia lemma. [http://en.wikipedia.org/wiki/Data_integration]

[2] A. Ramassotto et al. NRG4CAST D1.5, Early prototype of data gathering infrastructure, 2013.

[3] Y. Chamodrakas et al., NRG4CAST D2.4, Data distribution prototype, 2013.

[4] T.Hubina et al., NRG4CAST D1.6, Final prototype of data gathering infrastructure, 2014.

[5] T. Hubina et al. NRG4CAST D6.4, Real-time monitoring integration (2nd prototype), 2014.

[6] "Software Deployment", Wikipedia lemma. [http://en.wikipedia.org/wiki/Software_deployment]

[7] K. Kenda et al., NRG4CAST D2.3, Data cleaning and data fusion final prototype, 2013.

[8] K. Kenda, J. Škrbec and M. Škrjanc. Usage of the Kalman Filter for Data Cleaning of Sensor Data. In proceedings of IS (Information Society) 2013, Ljubljana, September 2013.

[9] A. Moraru et al., NRG4CAST D4.1, Complex Event Detection, 2014.

[10] K. Kenda et al., NRG4CAST D3.1, Modelling of the complex data space, 2014.

[11] J. Škrbec et al., NRG4CAST D3.2, Semantic Enrichment, 2014.

[12] K. Kenda et al., NRG4CAST D3.3, Metadata Generation, 2014.

[13] Web sourse: https://ucdavis.jira.com/wiki/display/IETP/About+CAS

[14] Web sourse: http://csrc.nist.gov/groups/SNS/rbac/

[15] T. Podobnik, T. Živko. On Probabilistic Parametric Inference. Journal of Statistical Planning and Inference vol. 142 issue 12. December, 2012, p. 3152-3166.

[16] M. Oliveira and L. Torgo. Ensembles for time series forecasting. Discovery Science, Bled, 2014.

[17] K. Kenda, L. Stopar, M. Grobelnik, Multilevel Approach to Sensor Streams Analysis, Discovery Science, Bled, October 2014.

[18] M. Bots et al. OGC Sensor ML: Model and XML Encoding Standard. 2014.