DELPHI: Data E-platform for Personalized Population Health

Yannis Katsis[†], Chaitanya Baru[§], Ted Chan[‡], Sanjoy Dasgupta[†], Claudiu Farcas[‡], William Griswold[†], Jeannie Huang[‡], Lucila Ohno-Machado[‡], Yannis Papakonstantinou[†], Fredric Raab[‡], Kevin Patrick[‡],

[†]Dept. of Computer Science and Engineering, [‡]School of Medicine, [§]San Diego Supercomputer Center
University of California, San Diego
La Jolla, CA 92093-0404
ikatsis@cs.ucsd.edu

***Abstract*—Recent studies recognize that health is influenced broadly by a multitude of factors of different types, including medical, genetic, environmental, social and behavioral factors. Developing successful health interventions therefore requires taking into account all these factors as well as the interactions between them. However, intervention designers have traditionally had access only to a very limited subset of health data (typically medical record data). Other health data, such as environmental or physical activity data, although already collected and stored, have been very difficult to access, since they are maintained by different providers and isolated in their own proprietary silos. This prevents physicians and intervention designers from acquiring a true overview of all factors influencing a condition and acting towards its prevention or cure.**

To solve this problem, we propose DELPHI: a platform allowing the integration of disparate health data into a single Whole Health Information Model (WHIM), providing a 360-degree view of an individual's health. DELPHI supports the integration of data and thus enables the design of applications and services that utilize the WHIM to offer the next generation of health services. In this paper, we describe DELPHI's architecture, outline the technical challenges encountered and describe an asthma management use case that will be enabled by DELPHI.

I. Introduction

Health and wellbeing are influenced by genetic, biological, medical, behavioral, social and environmental factors. In the case of asthma, determining factors of disease severity and exacerbation are known to be a combination of one's medical history and the environmental factors to which one is exposed (e.g., air quality). Likewise, obesity is influenced by an individual's medical history as well as behavioral factors, including diet and physical activity and environmental factors that either hinder or facilitate these behaviors. Similar observations can be made for most health conditions: The factors influencing health and disease are many, distributed across different levels of influence that interact with each other continuously and in subtle ways [4].

Thus, in order to determine the causes of a disease and design successful interventions for its prevention and cure, one has to holistically look at all the factors that affect it. Isolating certain subsets of those factors and designing interventions based solely on them may ignore important influences on health [11]. Yet, interventions today, be they for an individual or a population, usually ignore this interplay between different factors and focus on individual determinants instead. For instance, when designing a therapy for an asthma patient, a physician rarely takes into account the air quality in the places where the patient works or lives, since the physician does not have easy access to these data. Similarly, population-level interventions are still based on relatively generic (and hence not very representative) populations, such as "adult women", that may ignore other important factors, such as the environment in which an individual lives.

There is a simple explanation for the discrepancy between ideal (which calls for a study addressing multiple health-related factors) and practice (which focuses most of the time only on a small subset of these factors). Intervention designers, be they public health officials or physicians prescribing care to an individual, do not have access to all the data affecting an individual's health. For example, most of the time physicians only have access to their patients' health records localized to a single health system (i.e., not necessarily all their health information but only a part). The remaining health data are out of reach, as they are stored and maintained by different entities in their own isolated and proprietary silos. For instance, electronic medical record (EMR) data on obesity are maintained by health providers and insurance companies and are completely isolated from data on obesity correlates such as physical activity, now collected by an increasing number of wearable fitness sensors or smartphone apps (such as Nike+ or Fitbit) or data on diet increasingly captured today by mobile applications that help people self monitor calorie intake.

Thus, even though the health data exist, there is no easy way for a health practitioner to access them and consider them in unison to design successful interventions. In this work, we present DELPHI (Data E-platform Leveraged for Patient Empowerment and Population Health Improvement); a platform whose goal is to enable the integration of health data from disparate sources into a common data model, which can then be considered by health professionals to acquire an overview of all factors affecting a particular condition and use it as the basis for the design of interventions.

This work has been supported by the NSF grant “Information & Intelligent Systems, 1237174”.

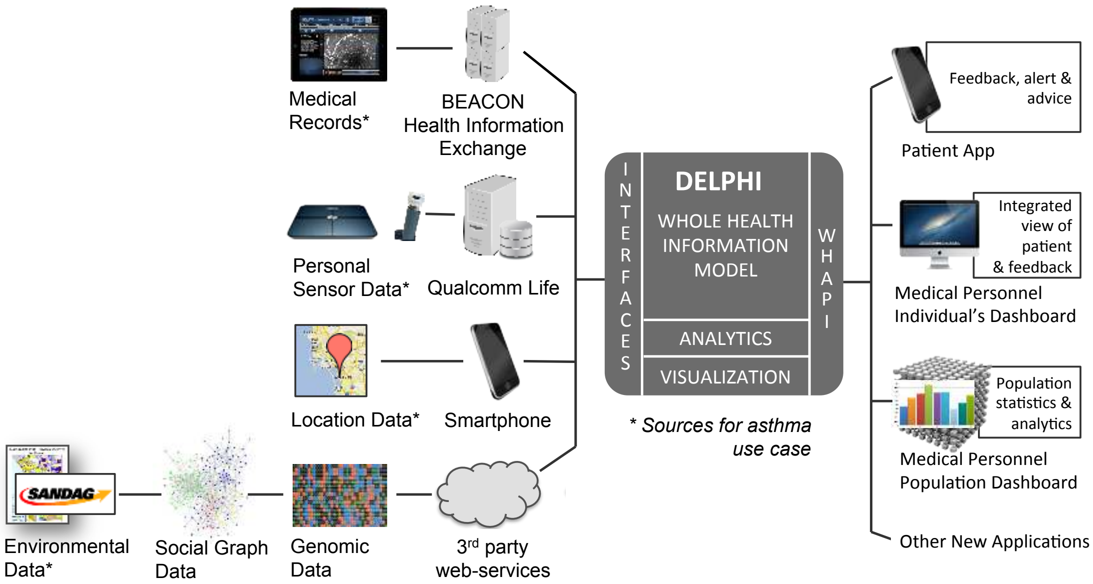

Figure 1. DELPHI Architecture

DELPHI, offered as a service to the health community, offers two main functionalities: (a) the integration of health data of different types into a single Whole Health Information Model (WHIM) and (b) an interface for accessing the integrated data to develop novel health applications that leverage the comprehensive view of an individual's health captured by WHIM to provide application users (e.g., medical personnel, patients, analysts or population-level intervention designers) with high-quality data. DELPHI's goal is to enable a new ecosystem for the development of services, where data owners on one hand and application developers on the other meet to create the next generation of health applications that are aware of the multifaceted nature of health.

This paper is structured as follows: We start by describing DELPHI's high-level architecture in Section II, related work in Section III and the novel research challenges that have to be addressed in Section IV. In Section V we describe a use case that will be enabled by DELPHI, showcasing both DELPHI's functionality and our vision for a new health ecosystem and finally we conclude in Section VI. As this is still work in progress, we do not delve into a detailed explanation of DELPHI's internals. Instead we focus on the system's high-level architecture together with the use cases that it will enable.

II. System Architecture

DELPHI's high-level architecture is depicted in **Figure 1**. Health data of disparate data types (shown on the left side of the Figure) are extracted from the data sources in which they reside and combined into the Whole Health Information Model (WHIM). This combined information creates a 360-degree view of an individual's health. Subsequently, different types of applications access these integrated data to provide specific information to its target users, be they patients, medical personnel, analysts or other people interested in health data.

To offer this functionality, DELPHI relies on five core elements: (a) The *Whole Health Information Model (WHIM)*; an extensible data model designed to capture all health-related data, which forms the heart of the system and interfaces with two components, namely (b) the *source interfaces*, allowing a source owner to register (i.e., add) a data source into the system and provide its data to the WHIM and (c) the *Whole Health API (WHAPI)*, allowing application developers to access the WHIM through their applications. Finally, to facilitate quick deployment of applications, DELPHI brings forward two data-oriented service layers for application developers. The first is (d) an *Analytics Layer*, allowing application developers to easily add efficient statistical computation capabilities into their applications and the second (e) a *Visualization Layer*, allowing them to enrich their applications with simple visualizations. We next describe each of these components in detail. As the system is still under development, the following discussion focuses on the functionality of each component and not on its implementation. We discuss the research challenges that arise in the context of DELPHI in Section IV.

Whole Health Information Model (WHIM): The Whole Health Information Model captures all the data that exist in the system (although this does not mean that all data are centrally stored in DELPHI, as we will explain shortly). It consists of a database schema that contains definitions for many common types of health data, such as environmental data, EMR data and fitness data. To accommodate new types of health data that might appear in the future, DELPHI offers schema evolution mechanisms that allow the addition of new concepts to the WHIM or the update of existing ones.

Source Interfaces: To bring health data into the WHIM, one can utilize DELPHI's source interfaces. These enable source owners to map their sources to the WHIM. To register a source its owner has to provide the following information: First, he has to specify which attributes of the WHIM his data source provides and second, the desired mechanism for retrieving and transferring the data. To create a flexible infrastructure that supports many different kinds of sources, DELPHI supports the following data access mechanisms: data streaming (useful for sensors collecting real-time information, such as heart rate monitors or fitness sensors), data copy (enabling the addition of information that is not accessible online, such as data contained in a spreadsheet residing on the source owner's laptop) and on-demand data access (used for sources that expose their content through web services). It is important to note that although DELPHI provides a single point of access to all integrated data, this does not translate to a centralized store. A subset of the data might be stored in the original sources (as is the case for on-demand access). In this case DELPHI employs caching to create local copies of frequently accessed data and thus improve its response time.

Whole Health API (WHAPI): Once the source data are registered in DELPHI, application developers can access them from their applications in an integrated form through the Whole Health API. The Whole Health API offers a query language allowing the expression of queries that retrieve the desired subset of the integrated data. Similarly to source interfaces, the application developer can specify whether he wants to retrieve the result once or continuously in a streaming fashion (useful for real-time monitoring and alert applications).

Analytics Layer: Given the substantial amount of data in the system and the data exploration nature of health research, a substantial number of DELPHI applications are expected to involve the execution of data-mining, statistics and machine learning algorithms on the integrated data. To ease the development of such applications, DELPHI implements an analytics layer that provides developers with a set of standard analytics functions they can invoke from their applications. The advantages of the analytics layer are twofold: On one hand, it accelerates application development by allowing developers to invoke the pre-implemented functions instead of implementing them from scratch. On the other hand, it improves efficiency, by leveraging optimization opportunities that are not available to applications. In particular, by being implemented in the same system that stores the WHIM, the analytics layer can push computation of the functions inside this system, instead of having first to extract all the data and then run the analytics on the extracted data. This works in the same spirit as the mechanism introduced in MADlib [9] for executing analytics inside a database management system.

Visualization Layer: The visualization layer is the analogous of the analytics layer for visualization functions, providing application developers with reusable modules for visualization. Thus, the benefits gained from visualization layer are similar to those obtained by the analytics layer.

III. RELATED WORK

Combining data from different sources is a long studied problem, known as *data integration*, that has led to a substantial amount of research [7] and to many industrial systems, commonly referred to as *Enterprise Information Integration* (in short *EII*) systems [6]. However, existing EII approaches are not directly applicable to Health Information Integration (HII) for two reasons: First, while EII has mostly focused on alphanumeric data, HII deals with additional data types, as for instance spatiotemporal or genomic. Although there has been work on databases for each of these data types [1][2], work on integration of such data is still in its early stages. More importantly we are not aware of any work that integrates in a principled fashion data *across* these data types, which is the goal of DELPHI. Second, in EII data sources and applications remain mostly static in the short term. This has led to EII solutions that require from system administrators considerable time and energy for source registration and data formatting so that queries can be evaluated efficiently [6]. In contrast, in HII new sources and queries appear continuously. These differences from EII create novel challenges for HII which we outline in the next section. Integration is also considered in Enterprise Application Integration (EAI). However, EAI focuses on integration of applications and not of data of novel types and is thus orthogonal to our work on data integration. Finally, taking into account additional context data has been studied in *context-aware systems* [3][12]. However, these works focus mostly on networking infrastructure or software engineering patterns to support *uninterpreted* context data and not on integrating *interpreted* data of particular data types, which is the main focus of this work.

IV. RESEARCH CHALLENGES

As discussed above, DELPHI differs from EII in that it needs to (a) consider new data types and (b) work in dynamic environments. This creates novel challenges, outlined below.

A. Challenges that arise because of the new data types

EII solutions have focused on alphanumeric and generally accurate enterprise-originating data. The targeted health data on the other hand have novel data types and varying degrees of resolution (e.g., county-level or street-level data) and accuracy (e.g., coming from medical devices or recreational-level sensors). To account for the different nature of health data types, we need to extend existing data integration architectures with support for new data types. This involves among others the design of a new data model that allows not only the representation of each of these novel data types in isolation but also the representation of a combination thereof (e.g., a combination of spatiotemporal and alpha-numeric data). Moreover, we have to design a query language for this model. Due to the importance of analytics and data mining in health applications, this language has to support statistics package (e.g, SAS) functionalities and analysis of the new types of data. Moreover, due to the varying accuracy of health data, the query infrastructure should also provide provenance information, identifying the origin of the data used in an analysis. This is especially important for doctors or regional planners making decisions based on the output of the query answering component. Last but not least, both the data model and the query language should provide adequate support for natural language annotations that are prevalent in the health domain.

B. Challenges that arise because of the dynamic environment

EII-oriented products are built on the assumption that the two major components in a data integration system - namely sources and application queries - change only rarely. Therefore it is considered acceptable to require from administrators to spend a considerable amount of time to accommodate both of these components: On one hand, to craft an exact mapping of the source to the integrated database and on the other, to reformat the data in ways that support the efficient execution of the application queries. The latter is typically accomplished through the design and implementation of OLAP data cubes; pre-computed views of the data that support efficient aggregation across a set of pre-defined axes [5].

In DELPHI, however, it is anticipated that data sources will be continuously added and that application requirements will change as developers create novel health applications that were not envisioned when DELPHI was designed. Therefore, one cannot afford to spend time registering sources or pre-formatting the data based on the application requirements. The source registration mechanism thus has to support the quick specification of source mappings. Similarly, query answering should gurantee the efficient evaluation of queries even in the absence of precomputed information. One technique to achieve this is approximate query computation. Based on the observation that users are often willing to trade accuracy for response time, especially for statistical queries, we will investigate novel techniques to compute approximate query answers, in the spirit of online aggregation [8].

V. Use Case: Asthma Management in the San Diego County

To demonstrate the feasibility of the approach and measure its effectiveness, we are using DELPHI to develop an end-to-end solution supporting two use cases. The first will help doctors manage and support patients with asthma, while the second will aid them in the prevention and cure of diabetes. To develop these use cases, we have formed a strategic regional alliance in the greater San Diego area, with participation of different entities that are currently maintaining different types of health data of interest. Below we describe in detail the asthma management use case. The goal of this description is twofold: First, to showcase a sample scenario that will be enabled by DELPHI and second, to demonstrate the novel health ecosystem that we envision will be built around DELPHI. While we focus next on the asthma use case, note that DELPHI is designed not for a single scenario, but as a general framework that allows the integration of multi-layered data to enable multiple health-related applications.

A. Data Sources

As studies have shown, the condition of an asthma patient is affected among others by his medical history, environmental factors, such as the quality of the air to which he is exposed, medication adherence and health behaviors, such as whether medication is being taken as prescribed, and exercise routines. Thus, in our use case we will be integrating the following types of data (denoted in **Figure 1** through an asterisk):

Electronic Medical Records (EMR): Patients' EMRs encompass all data collected during an individual's visits at a physician's office or hospital, including laboratory results and notes taken by doctors. For our use case, these data will be provided into DELPHI by the San Diego Beacon Health Information Exchange (HIE)[1]; an electronic network between healthcare providers, clinics, hospitals, emergency medical services and public health organizations that allows doctors to see patient health data from participating sites. San Diego Beacon will provide an API to its HIE platform, allowing individuals' medical data to be accessed through DELPHI.

Environmental Data: Environmental factors affecting asthma are known to be air quality and allergens present in the air. For our use case, we will acquire these data from two different sources, corresponding to different granularities. The first source is the San Diego Association of Governments (SANDAG)[2] and the San Diego County Department of Health and Human Services (SD HHSA)[3]. These governmental agencies collect GIS and air quality data for the greater San Diego area. Although not extremely detailed for a particular location, the provided data cover most part of San Diego county as sensors are located across the county. Our second source will be air quality sensors that individuals can carry with them, which were developed as part of the CitiSense project [10]. In contrast to the data acquired through SANDAG and SD HHSA, CitiSense data do not cover a large area, but rather focus on the path followed by an individual carrying the sensor and thus are much more detailed than the former.

Behavioral Data: To capture behavioral data associated to medication adherence, we will be interfacing DELPHI with the Asthmapolis[4] inhaler sensor; a sensor attached to a metered dose medication inhaler, monitoring when the patient uses his rescue medication when he is having difficulty breathing.

Activity Data: Activity data will be captured and transmitted to DELPHI through popular wearable sensors or smartphone applications, such as Fitbit[5] and Nike+[6]. To capture these sensor data, we will be utilizing Qualcomm Life's 2net platform[7]. 2net is a platform developed by Qualcomm Life, allowing medical and other health-related devices to transmit health data to servers. By interfacing DELPHI with 2net, we will leverage the Qualcomm Life's infrastructure and connect directly to devices that are already interfacing to 2net. This includes among others Fitbit, Withings (providing weight data), and the Asthmapolis sensor described above.

B. Applications

Once we have registered the data sources outlined above into DELPHI, developers will utilize the WHAPI to access the integrated data and develop the following three applications:

Asthma Patient Smartphone Application: Leveraging the data integration capabilities of DELPHI, asthma patients will have access to a smartphone application that automatically keeps track of data related to all major factors affecting their condition and uses such data to either alert the user of upcoming symptoms or record the data to be later interpreted by his physician. Figure 2 depicts two screens of a sample asthma patient application. The leftmost image shows an

[1] http://www.sandiegobeacon.org/

[2] http://www.sandag.org/

[3] http://www.sdcounty.ca.gov/hhsa/

[4] http://asthmapolis.com/

[5] http://www.fitbit.com/

[6] http://nikeplus.nike.com/plus/products/gps_app/

[7] http://www.qualcommlife.com/wireless-health/

overview screen, explaining at a glance the status of each of the factors affecting asthma. Through this screen, the patient can quickly see the air quality and allergens count at his current location, the status of his medication (whether it is overdue or whether he is running out of medication, as reported by the connected inhaler) and his current activity levels (based on data from the connected fitness sensor/app). By selecting any of these factors, he can get additional information on it. For instance, selecting the air quality will lead to a new screen showing the air quality of nearby locations, allowing him to revise his travel plans accordingly. The rightmost image shows the screen when the patient uses the inhaler connected to the Asthmapolis sensor. In this case the application records the event together with the entire environmental and behavioral context of the patient (i.e., location, air quality and activity data) under which it occurred, so that it can be later examined by the patient's physician. A “call” button also allows the patient to contact his health services provider if needed.

Medical Personnel Individual's Dashboard: When the patient calls or visits his healthcare provider, medical personnel will be able to utilize a DELPHI-enabled application to access all data pertaining to this patient. These will include among others data explaining the context under which asthma events occurred, allowing thus the physician to create a better estimate of what negatively affects the particular patient's condition.

Medical Personnel Population Dashboard: Until now, we have described applications that pertain to an individual’s health. However, DELPHI will also enable population-level health applications. In the asthma use case, such an application will be a population dashboard for medical personnel, allowing them to carry out population-level studies, such as computing statistics for asthma patients and finding correlations between factors affecting asthma. To offer this functionality, the dashboard will be internally utilizing DELPHI's analytics layer.

To evaluate WHAPI in real-life environments and receive feedback for improving the API and the system, we will be delegating the above development tasks to existing developers interested in digital health. To get in contact with them, we will be leveraging our partnership with UCSD CONNECT[8]; a regional entrepreneurship program that, having a strong track record of successfully launching many software companies, links inventors and entrepreneurs with the resources they need for success. In the context of DELPHI, CONNECT will attract application developers interested in interfacing with the WHIM to develop health applications. This community of developers will not only help bootstrap the application development, but it will also help test the integrated model and framework.

VI. Conclusion

To conclude, health professionals, although interested in all health data affecting a certain condition, usually do not have access to it. DELPHI is designed to address this problem by providing an infrastructure that not only allows the integration of such data into a common model, but also facilitates the implementation of applications that leverage this model to provide the next generation of health services.

[8] http://connect.org

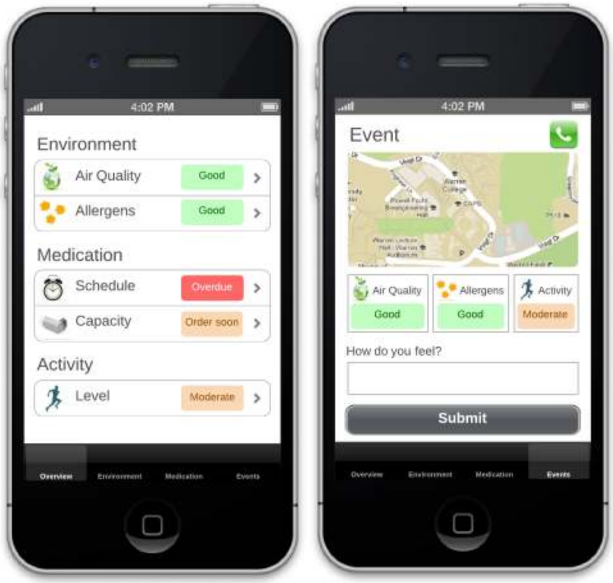

Figure 2. Sample Asthma Patient Smartphone Application

References

[1] T. Abraham, J. F. Roddick, “Survey of Spatio-Temporal Databases”, in *GeoInformatica*, vol. 3, no. 1, pp. 61-99, March 1999

[2] V. Bafna, A. Deutsch, A. Heiberg, C. Kozanitis, L. Ohno-Machado, and G. Varghese, “Abstractions for Genomics”, in *Communications of the ACM*, vol. 56, no. 1, pp. 83-93, Jan. 2013

[3] M. Baldauf, S. Dustdar, and F. Rosenberg, “A Survey of context-aware systems”, in *Int. Journal of Ad Hoc and Ubiquitous Computing*, vol. 2, no. 4, pp. 263-277, 2007

[4] T. A. Glass, M. J. McAtee, “Behavioral science at the crossroads in public health: Extending horizons, envisioning the future”, in *Social Science & Medicine*, vol. 62, no. 7, pp. 1650 – 1671, 2006

[5] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, M. Venkatrao, F. Pellow, and H. Pirahesh, “Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub-totals”, *Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 29–53, 1997

[6] A. Y. Halevy, N. Ashish, D. Bitton, M. Carey, D. Draper, J. Pollock, A. Rosenthal, and V. Sikka, “Enterprise information integration: successes, challenges and controversies”, in *Proceedings of the 2005 ACM SIGMOD*. New York, NY, USA: ACM, 2005, pp. 778–787

[7] A. Halevy, A. Rajaraman, and J. Ordille, “Data integration: the teenage years”, in *Proceedings of the 32nd international conference on Very large data bases*, ser. VLDB ’06. VLDB Endowment, 2006, pp. 9–16

[8] J. M. Hellerstein, P. J. Haas, and H. J. Wang, “Online aggregation”, in *Proceedings of the 1997 ACM SIGMOD*. ACM, 1997, pp. 171–182

[9] J. M. Hellerstein, C. Ré, F. Schoppmann, D. Z. Wang, E. Fratkin, A. Gorajek, K. S. Ng, C. Welton, X. Feng, K. Li, and A. Kumar, “The madlib analytics library: or mad skills, the sql”, in *Proc. VLDB Endow.*, vol. 5, no. 12, pp. 1700–1711, Aug. 2012

[10] N. Nikzad, N. Verma, C. Ziftci, E. Bales, N. Quick, P. Zappi, K. Patrick, S. Dasgupta, I. Krueger, T. v. Rosing, and W. G. Griswold, “Citisense: improving geospatial environmental assessment of air quality using a wireless personal exposure monitoring system”, in *Proceedings of the conference on Wireless Health*, ACM, 2012, pp. 11:1–11:8

[11] B. M. Popkin, K. Duffey, and P. Gordon-Larsen, “Environmental influences on food choice, physical activity and energy balance”, in *Physiology & Behavior*, vol. 86, no. 5, pp. 603 – 613, 2005

[12] Paweł Świątek, Krzysztof Juszczyszyn, Krzysztof Brzostowski, Jarosław Drapała, and Adam Grzech, “Supporting Content, Context and User Awareness in Future Internet Applications”, in *The Future Internet.* LNCS, vol. 7281, pp. 154-165, 2012