Unlocking the Value of Open Data with a Process-based Information Platform

Xavi Masip-Bruin[1], Guang-Jie Ren[2], Rene Serral-Gracià[1], Marcelo Yannuzzi[1]
[1] Advanced Network Architectures Lab (CRAAX), Technical University of Catalonia, (UPC), Spain
[2] IBM Almaden Research Center, San Jose, California, USA

Abstract- **There has been a wide shift in the way data are managed in the public administration. The move has led to an increased adoption of the Open Data model, where public administrations freely and openly publish data gathered using citizen taxes. However, undesirable side effects include the lack of data quality, incompatible formats and access methods, and various semantic interpretations of data. As a consequence, Open Data stakeholders, such as application developers, common citizens and even government agencies themselves, are overwhelmed by the large quantity of unstructured data, unable offer citizens and business value-added applications and services. To address the issue and make Open Data actionable, this paper proposes a systematic value-creation process that helps stakeholders identify the most suitable information assets and convert them into forms that can be more consumable by users. The process is enabled by the Middleware for Open-Data Aggregation (MODA), a platform designed with four main features, i) data quality assessment, ii) data homogenization for uniform access through an universal interface, iii) data correlation and semantic adaptation, and iv) secure data access. These features maximize the return on investment in Open Data by reducing time and cost of third party application development while providing improvement feedback to data sources.**

Keywords- **open data, value creation, data quality, data homogenization, semantic adaptation, third party developers**

I. Introduction and Motivation

Public organizations create, store and disseminate a great variety of information assets, ranging from demographic and economic to geographic and meteorological data. These data, also known as Public Sector Information, take different shapes and forms, located in different parts of the government. They can be raw data (e.g. population) or processed data (e.g. household income adjusted by cost of living). They may be related to public services (e.g. waiting time of an office visit for driver's license renewal) or internal processes (e.g. budgeting procedures).

Governments and public institutions across the globe are increasingly interested in the idea of Open Data [1], which is defined as "an approach to managing data so that it enables the structured free flow of non-sensitive information to those who have a need or interest in using it, both within and across government agencies and to the public. It allows different types of users to access, organize and use data in ways that make sense to them" [2]. In fact, a few public organizations, notably the US Federal Government, the UK Government (ca. 5,400 datasets as of April 2012), the World Bank and the City of San Francisco, have succeeded in initial efforts on Open Data [3].

Proponents of open data cite three rationales behind their support. First, open data makes government more transparent, participative and collaborative. With more information of what government is doing and how well it is performing, citizens as well as investors place more trust and interest in public affairs.

Second, open data encourages public involvement in data collection, analysis and application, often reducing government spending or improving efficiency accordingly. For instance, Washington DC engaged citizens to create software applications with the use of open data and in the first call they received 47 apps with a market value of $2.3 million, that is 50 times of the prize, not to mention broader social impact [4]. In a similar vein, New York City holds annual competition on web and mobile applications based on data available in the public. The winner applications are highly innovative and cost much less than if the government had decided to build on their own [5].

Third, open data creates a new source of economic growth. For example, the UK Central Government earns about 390 million British Pounds from direct supply of public sector information for commercial and personal uses [6]. European Union estimated a total of 140 billion Euros per year from open data, including direct contribution from data transactions and indirect contribution from information services [7].

Meanwhile, driven by both the advent of smart technologies and the wide connectivity offer, city councils and local governments are increasingly deploying different smart devices (sensors, detectors, etc.) and data gathering facilities all over our cities. This entire infrastructure, known as Smart Cities, turns into massive deployment of new community services, such as traffic lights optimization, efficient garbage recollection, public parks watering system optimization, or energy efficiency systems, all supported by an easy and wide connectivity offer, coining the Internet of Things (IoT). All these services provide a wealth of open data about the infrastructure and its usage. In fact, the combination of open data, wide devices deployment and pervasive connection capabilities all as a set, opens the door for a new scenario, where all stakeholders players may benefit from their combination, networks operators providing "smart" connection, services providers binding users to new emerging applications, third party software developers designing new applications, the users by utilizing new services (social, entertainment, etc.) and finally governments giving visibility to data that today is just collected but hardly used.

However, challenges of Open Data are abound, spanning from cultural and organizational to legal, skill and technological aspects. While Open Data initiative is a good starting point, it presents several issues that hinders adoption.

1. **Data selection:** the first issue is to decide what information assets are to be included in an Open Data initiative [8]. The prevailing approach is to start with 'low hanging fruits', information assets that are readily available for sharing without compromising public safety, security, and privacy. However, this “low hanging fruit” approach by nature is ad hoc. For one thing, it provides no comprehensive view of information assets inside and outside the organization. Furthermore, it is not clear if (and how) opening up the selected information assets contributes to the organization's vision and strategy. And very little if any consideration is given to the constraints in budget, resources and time.
2. **Data heterogeneity:** refers to the fact that the available data lacks a standard format and proper structure, as it ranges from structured semantic databases to plain text, spreadsheets, PDF files, among many others.
3. **Non-uniform data access:** refers to the fact that there is no uniform access method to Open Data, including web services, direct HTTP transfers, FTP downloads, and so on.
4. **Data security:** refers to the fact that there are no security guaranties to information access.
5. **Data quality:** deals with the parameters to be used to categorize the quality for a set of data . This information will be key to decide on both what set of apps may be developed based on the existing data and what quality is required to develop a particular set of apps. Quality assessment might be also useful to link data and apps according to certain quality parameters.
6. **Data processing:** referred to data interaction this step is mandatory to develop applications and services going beyond the simple exposition of data to users, what would significantly impact on fueling new business and market models.

These six issues pose severe limitations for third party application developers who today must individually customize collected data as well as its processing before turning any idea into a deployable application. As such, the main contribution of this paper is to propose a new framework, bringing together the various parameters and stakeholders to work together and facilitate the explosion of novel applications and services, fueling new revenue models impacting not only on the economical side but also on such aspects as social and health.

The paper is organized as follows. Section II provides an overview of the scenario, highlighting expected functionalities and open issues. Then Section III lists some references active in the data quality area, before introducing, in Section IV, the conceptual approach. Section V and VI then offers details of the Middleware for Open-Data Aggregation (MODA) and explains how it supports the adoption of open data. Finally, Section VI concludes the paper.

II. The Overall Picture

Fig. 1 depicts the overall system with main components, including performance features, inputs and expected outputs as well as a potential architectural location for the six open issues described in the previous section.

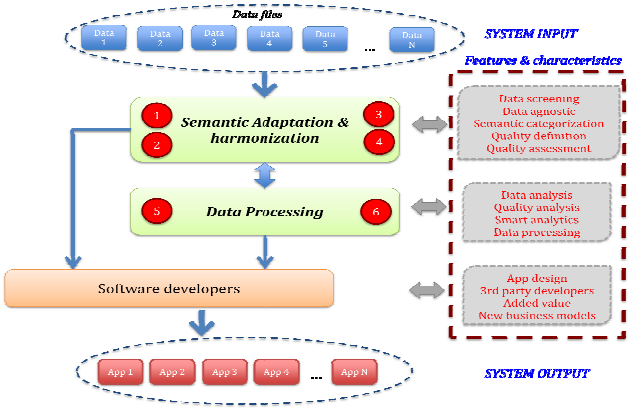

Figure 1. System zoom-out

The overall system feads from different heterogeneous (open) data files and ends up producing several applications and services, built by third party developers, and openly offered to users. It seems evident that the more "processed" the data is, the more "powerful" the applications might be. In fact, although different applications may come up from each individual data source, the added value resides in a more "elaborated" process (data interaction, correlation, etc.) supported by smartly mixing the content of different data sources. Therefore, software developers may build their applications either from "rough" isolated data or from data obtained after a certain data processing. In any case, the data coming from the heterogeneous sources must be harmonized to prevent third party developers to invest time on individual per-app/per developer customization.

The harmonization process can encompass several actions, from a simple format harmonization to a semantic categorization and adaptation (Semantic Adaptation and Harmonization in Fig.1). This block would be responsible for defining the quality parameters, mandatory to guarantee that apps built on top have the required data information. Once the data is harmonized the smartness is introduced in a second block (Data Processing in Fig.1), responsible for linking data quality parameters to apps and for handling smart analytics strategies .

In summary, the six issues mentioned in previous section may be mapped into two categories, data access (including data selection, heterogeneity and security) and data processing (including quality handling and smart analytics).

III. Insigts on Recent Contributions

Data quality is the most critical concept when dealing with data processing. In fact, it has been already introduced in [9], [10] that poor data quality might cause irreparable effects to economy and society. But, what does poor stand for?

Three strategies have been defined to dimension data quality [11], compared in Table 1. The first, intuitive approach, relates to experience of researchers. The second, theoretical approach, is highly dependent on the relative ratio of data quality during the data manufacturing process. The third, empirical approach, introduces correlation between consumers and data quality.

Table 1. Data quality approaches

Approach	***Advantages***	***Disadvantages***
Intuitive	Includes more relevant attributes	Do not consider consumers' feedback
Theoretical	Provides a comprehensive set of attributes	
Empirical	Concentrates on the consumers' feedback	Results cannot be validated

Extensive research has been conducted on data or information quality, in particular in the 1990s [12-14]. Several attributes have been already identified to characterize data quality [15-17], compelling from those more commonly used, such as completeness, interpretability, relevance, timeliness and accessibility to other not so used, such as clarity, comparability, conciseness, consistency, cost-effectiveness, easy of manipulation, traceability and usefulness.

Quality attributes play a key role to ensure existing data meets the particular specifications required by a concrete application to appropriately run. In other words, quality attributes may be utilized to select the set of applications to be developed and also to select the data meeting the expected quality demands for a particular application.

The concept in which the information is used, and the information consumers using the information are highly critical in the quality assessment process as they will help to determine the relevancy of the quality attributes, as well as the current and future quality level [18]. For instance, while a 20% error rate in customer addresses may be perfectly fine for telemarketing purposes, the same error is totally unacceptable for billing or collection services. It is therefore necessary to analyze the requirements on data quality from the viewpoint of data consumers that are most critical for the organization.

IV. The Conceptual Approach: Benefits & Objectives

1. The Concept

This paper proposes an architectural reference model that addresses the six issues previously identified and provides tools for linking data quality to the suitable applications and services to be deployed. This architectural model is conceptually supported by a platform named Middleware for Open-Data Aggregation (MODA), which takes a systematic and focused approach to establishing an Open Data initiative, mainly emphasizing on the following key points:

- MODA identifies information assets that create the best value for stakeholders with the lowest cost of data quality.
- MODA abstracts the access to heterogeneous sources of information. The effect is the availability of "*Open Data as a Service" (ODaaS),* where customers can access information without considering its origin or access method.
- MODA enables third party institutions to use a uniform RDF and Web Service based interface to access aggregated information from various data sources, with transparent guaranties about its origin and access rights, while using secure channels.

2. Benefits

The adoption of such a technology will deliver many-fold advantages to the current Open Data ecosystem, both from the technical and from the economic/social perspective.

Regarding the technical aspects, MODA offers the following key enablers all bundled together in a secure, easy to access, and extensible framework:

- selection of data with optimal stakeholder values
- technology independent uniform data access
- seamless integration of new services, and
- cross-data interaction and aggregation.

While technical aspects are essential in the MODA architecture design, another key goal is to provide a strong social and economic impact over the regions adopting open data. While MODA by itself does not provide a direct impact, it greatly simplifies the development of open data initiatives and thus increases the proliferation of novel Services and Apps requiring aggregated information from different heterogeneous sources. Moreover, such MODA enabled applications are likely to create an ecosystem able to make a difference to the cities, including: *i)* benefits from the selection of optimal open data to publish; *ii)* benefits obtained by third parties through the exploitation of such data, and; *iii)* social impact of a better infrastructure for the day-to-day life of the citizens.

3. Objectives

MODA aims at defining a framework assisting in the selection, orchestration, and integration of multiple external data sources into a single ecosystem. Consequently, MODA is focused on four high level objectives:

- *Data selection*: Focused on the evaluation and quality assessment of information assets in order to analyze its stakeholders and its suitability to be part of the potential set of Open Data offered by the administration. Then, rather than focusing on the "low-hanging-fruits"

information assets [19], we propose a structured and well-defined information assets quality evaluation process and the added value it presents to the society. Such data selection analysis will be separated into two different parts. First as an off-line process, which will determine the most relevant information assets required as Open Data. And second, an iterative on-line process that will periodically assess the quality of the available information within MODA and from the third party applications perspective.

- *Data acquisition*: MODA includes a data acquisition method featuring transparent protocol, transparent data format, and update frequency independence. This gathered information, will later be analyzed through a data processing subsystem**.**
- *Data processing:* Depending on applications, two invocation modes may be offered. On the one hand, MODA will provide a full-featured set of functionalities composed by smart multi-source data processing and semantic data correlation and adaptation mechanisms. On the other hand, depending on the requirements of the third party applications, the data processing will be outsourced and controlled by the external app. In this case, the application may use a reduced set of capabilities offered by MODA, for example multi-source data adaptation.
- *Data delivery*: The last objective is to provide a standard and flexible data delivery mechanism to the final app or service. MODA will be presented to third parties through a secure, easy to use, and well-defined Web-Service based interface.

To illustrate these objectives with a practical use case, Fig.2 shows the different steps in a MODA environment. In this example, a city council offers to the citizens a variety of data about public services, i.e., street issues, social events within the urban area, local newsletter, and map information. The particular information assets delivered as Open Data are refined and proposed by the data screening and selection process offered by MODA. The main goal of data selection is to maximize the revenue obtained by the published data.

In this example, the data published by the administration for each area has its own format and update frequency. For instance, the list of street issues is in a plain text file and the social events that are stored in a structured XML file. They have a different update frequency as they are manually entered on demand to the system, while the periodic newsletter is stored using a binary format as PDF, and has a weekly issue. Meanwhile, the urban map uses the common GIS format, and is static and it is rarely updated. Without MODA interface, any application willing to use such information would have to implement manually by interpreters for each data format and by schedulers to verify whether the data is up to date.

After the data acquisition as a decoupled process from the different data sources, the data collected is then processed and evaluated using semantic data information to assist in the data analysis features offered by the external applications accessing the service. In this case, in the absence of MODA, the apps developed would need a customized data processing engine, which would be developed for that particular purpose, resulting in added cost. In comparison, MODA takes over the query processing to the external applications.

Finally, all this information is made available through a secure Web Services. Then, a third party application, by having access to the API is able to perform queries over the information, that might be as simple as just asking for the list of street issues, or rather contain semantic information which requires data from different sources, such as asking about news regarding robberies during a time frame over a geo position.

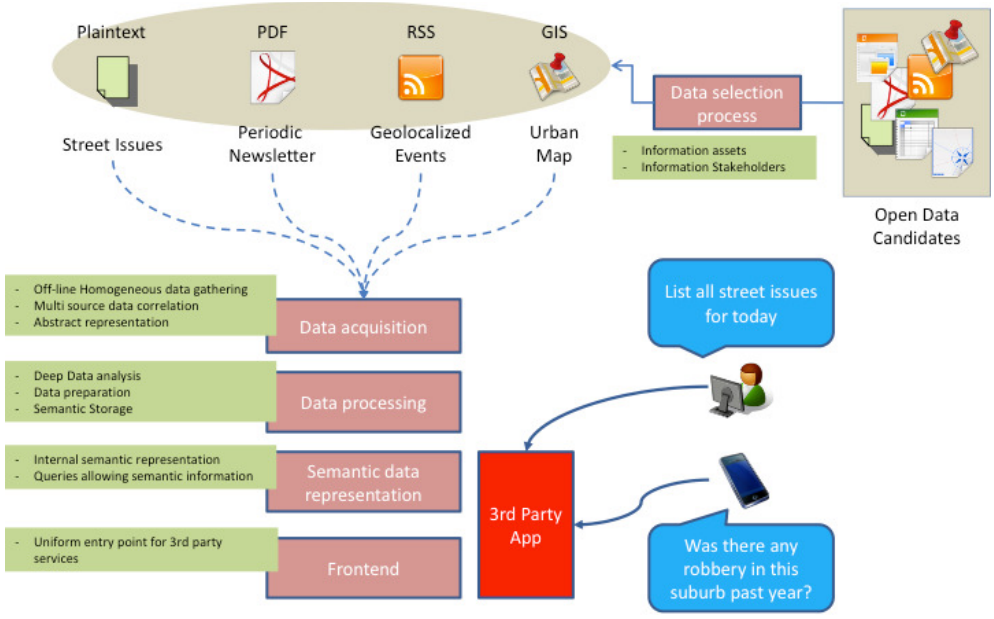

Figure 2. Practical example for a MODA use case

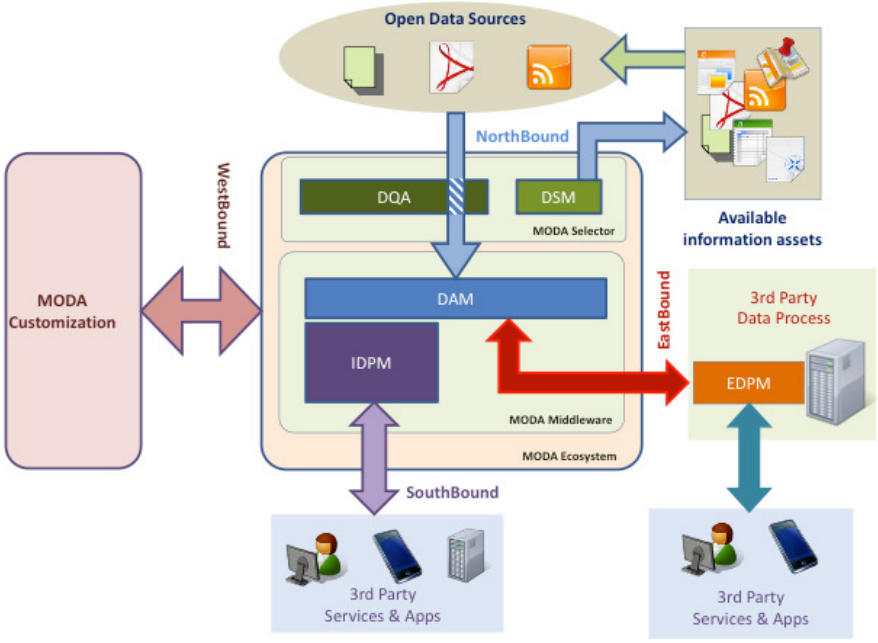

Figure 3. MODA Architectural Model

Without the presence of MODA, the third party application instead of directly invoking an easy to use service and hence focusing on the end-user interface, it would need to develop both the interface and the interaction with the gathered data. As such, MODA transforms all the above mentioned complexities into a neat and simple interface performing all the "under-the-hood" processing while keeping a high level of customization and availability for new services.

V. Architectural Model: The Functional Blocks

The reference model used in MODA is shown in Fig.3. The proposed architecture for the MODA ecosystem is mainly composed by three functional blocks, the MODA Selector, the MODA Middleware, and the MODA External Entities:

1. The MODA Selector

The MODA Selector evaluates the "quality" of the existing information assets to maximize the benefits of the Open Data stakeholders. The goal is to improve the set of available Open Data sources while analyzing the quality and proposing enhancements to data sources. More specifically, The MODA Selector is composed by two different functional modules:

- *Data Selector Module (DSM):* The DSM is centered in the research of novel Open Data assets in order to improve the quality of the overall portfolio. Then, rather than uniquely relying on the existing Open Data assets, MODA Selector considers actual requirements of the real apps, together with the availability of the Open Data sources to propose new data sources to the system.
- *Data Quality Assessment (DQA):* The DQA is centered in the evaluation of the available Open Data in order to provide mechanisms to improve its quality, filtering and screening the available data. Thus, proposing upgrade vectors and guidelines to the public administration about the type of data to release, maximizing the revenue of the whole Open Data assets syndication process plus the added value of the third party applications benefiting of such new information assets.

2. The MODA Middleware

The MODA Middleware complements the MODA Selector with two features. First, providing an easy-to-use data acquisition mechanism from different, heterogeneous data sources into a single generic representation for homogeneous access. Second, providing a secure, flexible, and intuitive framework for external third party applications and services to offer added value from those multi-source Open Data.

The MODA Middleware is composed by two internal modules, namely the Data Acquisition Module (DAM) and the Internal Data Processing Module (IDPM).

- *Data Acquisition Module (DAM):* The DAM, which interfaces through the Northbound with the external data sources, has the goal of providing data fetching and access abstraction to the rest of the MODA system. As shown in Fig.4, DAM is composed by the following entities:
  - *Fetch Agents (FA):* these pluggable agents are in charge of fetching the required data from the external data sources and pass it to Raw Data Cache Subsystem. FAs are designed in the form of small plug-ins, which are easily customized and deployed into the system using the EastBound Interface.

- *Raw Data Cache Subsystem (RDCS):* This subsystem has two different roles. First it gathers a copy of the data as fetched by the FAs, and second it provides high availability of the data to the rest of the system.
- *Data Abstraction Subsystem (DAS):* The DAS, by using a data description template provided by the data description database, or by information provided by the third party involved in the fetching, the raw data will be populated using standard formats such as RDF.
- *Data Correlation Subsystem (DCS):* This subsystem will provide a first level correlation among the different data sources, by providing a system for cross-data sources queries.
- *Data Access Interface (DAI):* Finally the DAI will provide a user friendly web-service like interface for the rest of the modules and third parties in order to access this gathered data.

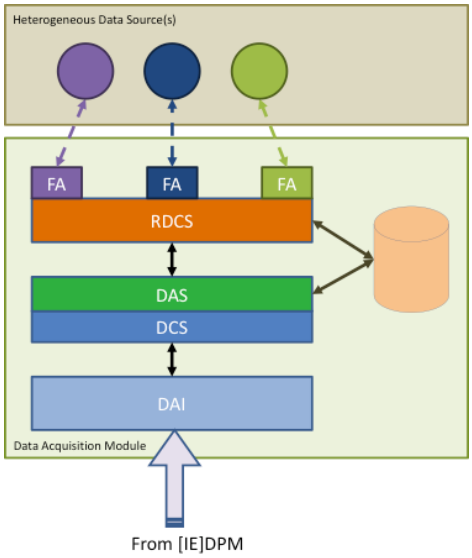

Figure 4. Data Acquisition Module Internal Architecture

- *Internal Data Processing Module (IDPM):* The main task of the IDPM is to provide required data processing capabilities to ease the task for third party applications. Particularly, this subsystem provides semantic data adaptation with ontology mapping for the different data sources, plus the possibility of outsourcing the high processing power required by such queries. It allows applications to be executed efficiently in all kind of devices, e.g., tablet, smartphones, workstations, and so on. As shown in Fig. 5, There are different parts composing the IPDM:
  - *Semantic Query Engine (SQE):* This engine is in charge of providing a smart query and a data filtering system able to combine the information acquired from the DAM by using ontology-like structures that allow the whole system to provide a high degree of semantic adaptation to the third party customers.
  - *Lightweight Portable Query Library (LPQL):* This library resides in the IDMP and provides simple yet powerful semantic primitives to the system. A particularly powerful capability of this library is that is designed to be deployed and used within in the third party App through the EDPM in the case of outsourcing the processing of the acquired data.
  - *Web Service Interface (WSI):* This provides the easy to use interface of the whole MODA system through Web Services.

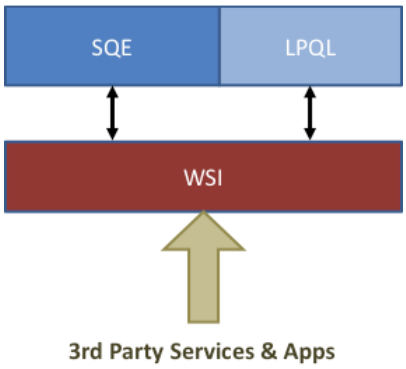

Figure 5. Internal Data Processing Module

3. The MODA External Entities

The MODA External Entities act as an external entity allowing third party interaction and users customization. The communication between MODA and the external actors within the ecosystem is performed through four different Interfaces:

- *NorthBound Interface* allows gathering data from the multiple available Open Data sources, while at the same time it is used to monitor and quantify the quality of the available Open Data assets.
- *EastBound and SouthBound Interfaces* are used by third party applications and services that benefit from MODA solely on the multi-source data quality assessment process and its homogeneous data acquisition interface, while relying on their internal data processing capabilities to finally offer the service. These two interfaces are devised for third party applications with sufficient internal infrastructure to offer the service, using MODA as a data abstraction platform.
- The *WestBound* Interface is designed to offer a MODA Customization facility, where the different actors within the MODA ecosystem are allowed to adapt the behavior of the platform to their particular requirements, in terms of data acquisition and data processing capabilities.

More specifically, there are two external entities allowing MODA to interact with users and third party entities/applications:

- *External Data Processing Module:* This module is not part of the MODA Middleware per se, as it composed by a subset of the internal IDMP libraries, which are available by the third party applications and services, in the case that such apps have enough resources to cope with the computational and time constraints present in the offered apps.
- *MODA Customization:* This module, not part of the MODA Middleware either, it allows users to customize the way they access the MODA system as well as the way the system handles the collected data, in order to make data process close to the expected users requirements.

VI. CONCLUSIONS

In this paper, we introduce a conceptual approach proposing an structured and systematic value-creation process that helps stakeholders identify the most suitable information assets and convert them into forms that can be easily consumed. The process is enabled by the Middleware for Open-Data Aggregation (MODA), a platform designed with four main features, *i)* data quality assessment, *ii)* data homogenization for uniform access through an universal interface, *iii)* data correlation and semantic adaptation, and *iv)* secure data access. The proposed process maximizes return on investment in Open Data by reducing time and cost of third party application development while providing improvement feedback to data sources.

The ideas presented in the paper are still work in progress and hence aiming at fueling further research to validate the approach in various contexts, create detailed guidelines and develop an assessment tool. We plan to introduce and test the process in selected municipalities, developing several real use cases scenarios.

ACKNOWLEDGEMENTS

This work was supported in part by the Spanish Ministry of Economy and Competiveness under contract TEC2012-34682, and the Catalan Research Council (CIRIT) under contract 2009 SGR1508.

REFERENCES

[1] K. Lakhani, R. D. Austin, and Y. Yi, "Data.gov," Harvard Business School Case Study Serial Number 9-610-075, 2010, USA.

[2] J. W. Cortada, V. A. Nix and L. C. Reyes, "Opening up government: How to unleash the power of information for new economic growth," IBM Institute for Business Value, 2011, USA.

[3] B. Hogge, "Open Data Study: New Technologies," Open Society Foundation, 2010, UK.

[4] "Apps for Democracy", http://www.appsfordemocracy.org/

[5] NYC, http://nycbigapps.com/

[6] D. Newbury, L. Bently, R. Pollock, "Models of Public Sector Information Provision via Trading Funds," UK Department of Business, Enterprise and Regulatory Reform, 2008, UK.

[7] G. Vickery, "Review of recent studies on PSI re-use and related market developments," Information Economics, 2011, France.

[8] G. Ren, S. Glissmann and J. Sanz, "Identifying Information Assets for Open Data", IEEE Conference on Commerce and Enterprise Computing, Hangzhou, China, September 2012.

[9] R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers", Journal of management information systems, pp. 5-33, 1996.

[10] T.C.Redmann, "The Impact of Poor Data quality on the typical Enterprise", Communications of the ACM, 41, 2, February 1998

[11] D. M. Strong, et al., "Data quality in context," Communications of the ACM, vol. 40, pp. 103-110, 1997.

[12] R.Y.Wang, D.M.Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers", Journal of Management Infoprmation Systems, 12, 4, 1996

[13] R.Y.Wang, "A Product Perspective on Total Data Quality Management", Communications of the ACM, 41, 2, February 1998

[14] Y. Wand and R. Y. Wang, "Anchoring data quality dimensions in ontological foundations", Communications of the ACM, vol. 39, pp. 86-95, 1996.

[15] Y. W. Lee, et al., "AIMQ: a methodology for information quality assessment," Information & Management, vol. 40, pp. 133-146, 2002.

[16] L.L.Pipino, Y.W.Lee, R.Y.Wang, "Data Quality Assessment", Communications of the ACM, 45, 4, April 2002

[17] A.Maydanchik, "Data Quality Assessment", Technics Publications, 2007

[18] J.E.Olson, "Data Quality: The Accuracy Dimension", Morgan Kaufmann Publishers, 2003

[19] G. Lee and Y. H. Kwak, "An Open Government Implementation Model: Moving to Increased Public Engagement," IBM Center for the Business of Government, 2011, USA.