View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Ghent University Academic Bibliography

Data Harvesting, Curation and Fusion Model to Support Public Service Recommendations for e-Governments

Gayane Sedrakyan[1], Laurens De Vocht[1], Juncal Alonso[2], Marisa Escalante[2], Leire Orue-Echevarria[2], Erik Mannens[1]
[1]*IMEC / IDLab, University of Ghent, Belgium*
[2] *Fundacion TECNALIA Research & Innovation, ICT – European Software Institute Division, Derio, Spain*
[1]*{name.surname}@ugent.be,* [2]*{name.surname}@tecnalia.com*

Keywords: Architectural model, recommendation generation, public administration, public services, data harvesting, data curation, data fusion, linked data, e-Government

Abstract: This work reports on early results from CITADEL project that aims at creating an ecosystem of best practices, tools, and recommendations to transform Public Administrations with more efficient, inclusive and citizen-centric services. The goal of the recommendations is to support Governments to find out why citizens stop using public services, and use this information to re-adjust provision to bring these citizens back in. Furthermore, it will help identifying why citizens are not using a given public service (due to affordability, accessibility, lack of knowledge, embarrassment, lack of interest, etc.) and, where appropriate, use this information to make public services more attractive, so they start using the services. While recommender systems can enhance experiences by providing targeted information, the entry barriers in terms of data acquisition are very high, often limiting recommender solutions to closed systems of user/context models. The main focus of this work is to provide an architectural model that allows harvesting data from various sources, curating datasets that originate from a multitude of formats and fusing them into semantically enhanced data that contain key performance indicators for the utility of e-Government services. The output can be further processed by analytics and/or recommender engines to suggest public service improvement needs.

1 INTRODUCTION

This work reports on early results from the CITADEL, a H2020 European project (CITADEL Consortium, 2017) that aims to create an ecosystem of best practices, tools, and recommendations to transform Public Administrations (PAs) with more efficient, inclusive and citizen-centric services. The CITADEL ecosystem aims to improve the processes and policies of the PAs using what they already know plus new data to implement what really matters to citizens in order to shape and co-create more efficient and inclusive public services. The innovative ecosystem that builds on the best practices innovates by using ICTs to find out why citizens stop using public services, and use this information to re-adjust provision to bring them back in. Also, it identifies why citizens are not using a given public service (due to affordability, accessibility, lack of knowledge, embarrassment, lack of interest, etc.) and, where appropriate, use this information to make public services more attractive, so they start using the services.

In this work we extend The DataTank (Vander Sande, 2012), to provide the Data Harvesting/Curation/Fusion (DHCF) component of the platform based on which recommendations for the utility and improvements of public services as well as suggestions for specific services will be generated to PAs. The DataTank provides an open source, open data platform which not only allows publishing datasets according to standardised DCAT-AP guidelines and taxonomies promoted by Open Data Support (http://opendatasupport.eu), but also transforms the data into a variety of reusable formats. This allows PAs to publish data in an almost effortless manner, with maximum impact in terms of visibility. Using this platform civil servants will see their open datasets automatically being

crawled by other aggregation portals (a.o., the EU open data portal) because of the DCAT-AP compliance. The extension will include an intelligent way of harvesting and fusion of different (big) data sources using semantics and Linked Data technologies. In the context of CITADEL the new DHCF component will enable the visualization and analysis of trends for the usage of public services in European cities, playing a key role in in terms of suggesting improvements to the current suite of public services. This will allow rising the PAs' knowledge regarding their progress across various e-government Key Performance Indicators (KPIs) to improve and make more specific and evidence based on their e-government investment plans (see KPI examples in the section on architectural design). In long term it would have a positive impact on more efficient and effective e-government investment strategies of public institutions.

While the approach that will be followed in CITADEL for the big data analysis is not novel, the CITADEL solution regarding big data algorithms innovation lies on 1) the domain (public sector) in which it will be applied, 2) the purpose for what is created, that is, the creation of KPI reports containing business intelligence that will be used as input to derivate generic (semi-)automatic recommendations to improve the processes and policies of the PAs. The focus of this work is to provide an architecture that will allow collecting data from various sources in different formats (e.g. e-Government portals, offline data, other online sources such as social media) and fuse them into a semantically enhanced dataset in order to facilitate more efficient and inclusive analytics and recommendation processes for PAs.

2 RELATED WORK

Recent R&D topics show increased interest in the use of recommender systems for e-Government to assist with customized suggestions for the use of public services. While recommender systems can enhance the user experiences by providing targeted information, the entry barriers in terms of data acquisition are very high (Heitmann, Hayes, 2010). To our knowledge scientific publications describing research-based approaches and methods for harvesting data from multiple sources, curating and combining different datasets as basis for recommendations in public service domain are largely lacking. Often, the scope of recommendations is also limited to user models and context variables that need to be constantly updated by human interpreter to consider new variables and maintain the semantics between different model variables. To the best of our knowledge, only one recommendation approach has been presented that focuses on the e-Government service recommendations that relies on semantic knowledge using semantic ontologies. Yet, the focus of the recommendations is limited to one specific are for tourism (Al-Hassan et al., 2015). In this paper we posit that harvesting of context variables and KPIs for visualizations for e-Government service recommendations can be extended to rely on open data that may exist beyond such models, e.g. anywhere from web (e.g. social media discussions) or European portal, which may be collected and transformed into unified dataset that is ready to be processed by recommender engines. We posit that linked data technologies will allow fulfilling this task in an automated way by also maintaining the semantics from different sources and formats.

The Semantic Web provides technologies for knowledge representation, which can deliver Linked Data created by multiple parties at Web scale. For any given entity in a recommendation database, the open world assumption means that we can harvest more contextual information by looking up data on the Web through link following. Because identification of concepts happens through universal identifiers - as opposed to local database IDs - other parties can attach additional metadata to any concept in order to improve recommendations.

3 ARCHITECTURAL DESIGN OF THE PROPOSED METHOD

The KPI visualization and Report generation component of the CITADEL ecosystem will generate a report based on filtered KPIs. The report will be presented as visualizations to support recommendations to PAs. The data will be checked for privacy sensitivity and anonymized if needed. The process flow and possible UI mockup for some KPI definitions/filters are shown in Figure 1 and Figure 2 (CITADEL Consortium, 2017) subsequently. Examples of possible KPIs include:

- KPIs to co-create: Number of users and trends
- General KPIs for improving the usage of the current digital services in general as well as for a specific service:
  - Number of users/non users per service/per year
  - Data/information of citizens who (do not) use the digital services (or one concrete

digital service): i.e. demographics such as age, gender, education and computer skills, internet access, devices used and frequencies of using digital devices, other methods and causes contacting governments (e.g. offline visits, phone calls), etc.

- Built-in feedback (rating) per service/ per type of service/ per year (e.g. problems using services, satisfaction, etc.)

These are the possible KPIs envisioned by now. The list of possible KPIs might be extended to include further input from CITADEL co-creation methodology and its subsequent component.

The Data Harvesting, Curation and Fusion component (DHCF) in the context of KPI visualizations will collect, store, fuse and provide data related to the specified KPIs. The main functionalities of this component are:

- Harvesting/loading structured file based resources in different formats (like CSV, XLS, JSON, XML, SQL, RDF…) but also databases through SQL or indexes like Elasticsearch and publishing data in different formats (like CSV, JSON, XML, RDF…).
- Filtering resources based on their metadata, for example based on keyword or resource type.
- Linking and fusing data sources by adding semantic context.
- User management: set the visibility of resources for certain resources according to user/groups.

The architecture of the DHCF component consists of three subcomponents: The DataTank (Vander Sande et al., 2012), RML Mapper (Dimou et al., 2014) and RDF Store (W3C, 2014). There are two main parts, which are able to work independently.

Part one focuses on loading, harvesting, managing, curating and (re)publishing structured data sources (mainly files) in different formats. Part two focuses on fusing the loaded/harvested sources together.

As it can be seen in the Figure 3 there is a central component the Resource Description Framework (W3C, 2014) (RDF) Store that it is shared by the two parts. The RDF Store allows storing and querying semantically linked graphs.

Introducing semantics to structured data is important because it allows explicitly indicating the context of the different data sources being combined. In RDF, The figure below gives an overview of the components of the proposed architecture. The main component for data collecting is the DataTank. The DataTank allows to harvest, load and manage (or curate) different structured data sources. It has built-in a SPARQL (W3C, 2013) templating mechanism to access linked data residing in an RDF Store, which is necessary for publishing combinations and fusions of the data sources. The main component for fusion relies on RML Mapper (Dimou et al., 2014), the objective of which is to convert separate sources, that may have different structures or be in a different format to linked data. By converting data sources to linked data, the main component RML Mapper, is to some extent also fusing the data or at least preparing the data in a linked format (RDF) so that they can be fused later, for example through a SPARQL query.

The data sources are accessed through the HTTP interface that is provided by the DataTank. The process followed by these components supports three main steps:

1. **Adding a data** source and publishing it through HTTP.
2. **Combining data**, with a SPARQL query to the RDF Store, and adding it as a combined 'fused' data source.
3. **Mapping** one or more data sources to linked data after they are exposed in the source interface and loading them into the RDF store.

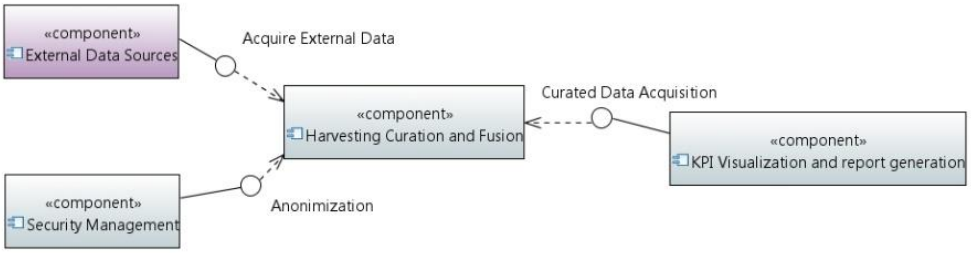

Figure 1: KPI visualization components in CITADEL.

Figure 2: Sample UI for KPI definitions.

In addition to these steps, the component will have two different behaviours:

- **Creating new resources.** There are different ways to create new resources:
  - Creating a new dataset from a single resource using SPARQL query. The user triggers the retrieval of the data source through the Web UI. The DataTank will create a reference to the data source and expose the data source via a HTTP interface.
  - Creating a new dataset from a combined resource using SPARQL query (manual query). The user triggers the execution of a SPARQL query after configuring the SPARQL query manually (currently supported) or via selecting the properties and original data sources to combine (to be implemented). Like with a single data source, the DataTank will create a reference to the data source and expose the data source via a HTTP interface.
  - Creating a new dataset from a single resource using SPARQL query (mapping semantics). Mapping/linking data sources is currently done through a mapping document.
  - Creating a new dataset from a single resource using SPARQL query (alternative automated solution).

The output path will follow the data naming convention specified in the DataTank documentation: path": "/definitions/{identifier}, where: {identifier} consists of 1 or more collection identifiers, followed by a final resource name. (e.g. world/demography/2013/seniors). The convention will also support the searchability of resources.
Furthermore, the DataTank allows to categorize resources under the following naming convention: http://example.org/{category}/{resourcename}. The metadata will be described in DCAT-AP format, profiting from the functionality provided by the DataTank.

- **Retrieving resources.** There is a distinction between retrieving a single data source and a combined data source.
  - **Single**: A request to retrieve a single data source is parsed through the DataTank's HTTP interface and results in processing the requested data source in the requested format and returning it to the application where the request originated from.
  - **Combined**: When retrieving a combined data source, the incoming request is translated to the configured SPARQL Query. It is the RDFStore that contains the linked data of all data sources that have been mapped. So retrieving a combined data source is only possible after mapping the data sources that need to be combined and specifying a SPARQL query (template) needed for the combination of data sources.

Finally, data can as well be destroyed as required.

The harvesting and curation component will be an internal sub-component of the KPI visualization component. This sub-component will interact with KPI visualization and report generation component to receive the request of the data, with the Security Management to request the anonymization / encryption of certain data and with external data sources to get the data (Figure 1).
When adding a combined resource, the DataTank generates an RML mapping document involving the selected data sources. The generated mapping document maps the source files according to the chosen and mapped properties in the data sources. The DataTank also generates a SPARQL query that selects the chosen output properties for this resource. The resulting combined data source will appear in the list of available resources. It will behave similar to a SPARQL resource. The main difference is that the creation of the mapping document and the SPARQL query will be hidden from the user.

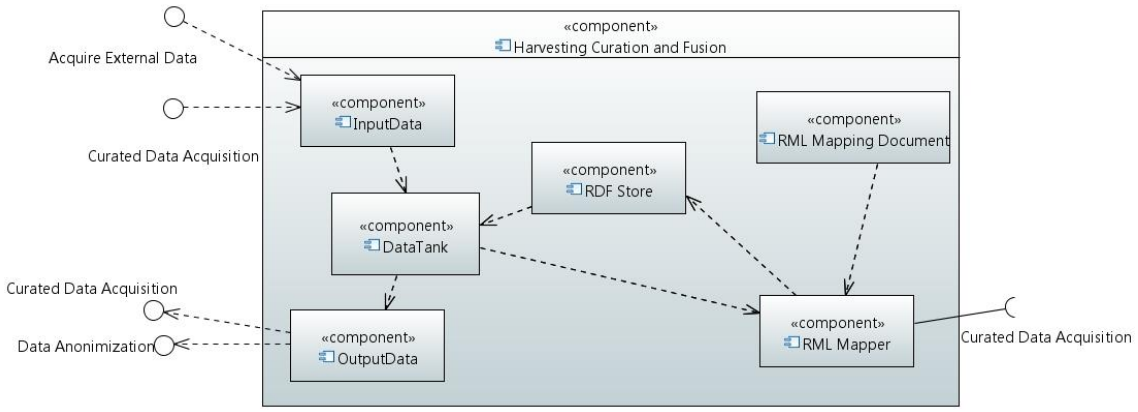

Figure 3: Harvesting/curation/fusion component architecture.

Currently the mapping documents and SPARQL queries need to be manually configured. Figure 4 shows an example of such mapping document. The SPARQL queries in the DataTank and the mapping documents are on a mounted local file system or file server. In the future, it will be possible to select data sources as well as the properties to fuse them on and also the target properties to map the source properties on. It is very common that a similar property might use a different column name to depict the same.

```
{
    column_mapping: [
        {  final_name: "TIME",
           columns: [
               "TIME"
           ]},
        {  final_name: "LOCATION",
           columns: [
               "GEO"
           ]},
        {  final_name: "TYPE",
           columns: [
               "INDIC_IS"
           ]},
        {  final_name: "INDIVIDUALS",
           columns: [
               "IND_TYPE"
           ]},
        {   final_name: "UNIT",
           columns: [
               "UNIT"
           ]}
    ],
    filters: {
        TIME: {max: 2006}
    },
    sort: {id_column: "asc"}
}
```

Figure 4: Mapping and filtering example with manual configuration file in JSON format.

To add a combined resource, it is necessary to formulate a SPARQL query. RML enables reusing mapping definitions/configurations to be used with different formats (Figure 5).

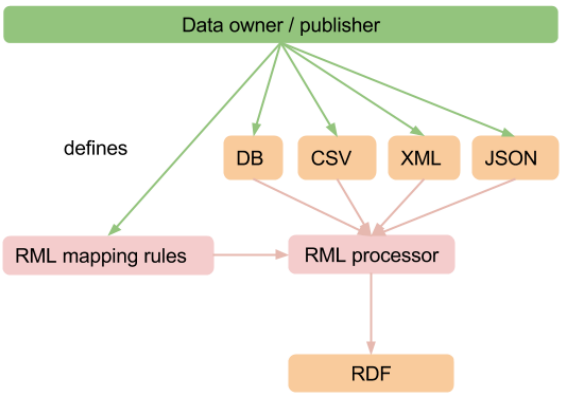

Figure 5: Fusing different sources using RML mapping definitions (Dimou et al., 2014).

Furthermore, RML provides a solution to model domain-level knowledge in a scalable, integrable and interoperable fashion by semantically representing data derived from multiple heterogeneous sources using the RDF framework. RML uses uniform mapping definitions that are independent of the references to the input data. RML mapping definitions (Dimou et al., 2014) are 1) reusable across different sources; 2) interoperable across different implementations for different source formats which allows reusing them at reduced implementation and learning costs; 3) scalable and extendable allowing to reference to the data extracts and the mapping definitions in a distinct fashion by the use of generic way of definitions for what can be

used for all possible different input sources and scales over what cannot.

To avoid compatibility issues as well as to facilitate versioning upgrades, the initial version of the DHCF will rely on an interface that will link functionalities from the DataTank and RMLMapper using REST services. Figures 6-10 show examples of the user interfaces for the initial version of the DHCF component. Figure 6 shows the harvesting interface mockup that will allow easy loading of datasets from sources such as local files, files residing on web, database tables as queried data. Figure 7 shows the content panel interface where the loaded resources can be explored and altered.

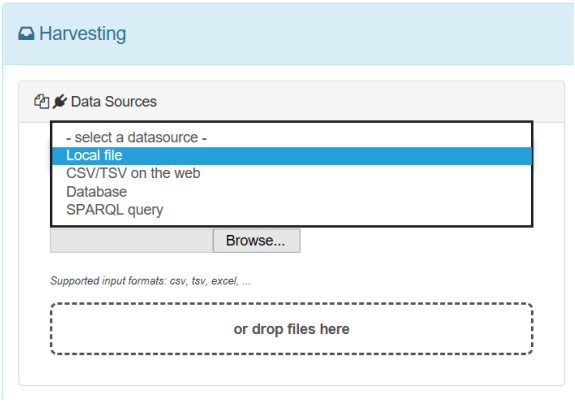

Figure 6: Harvesting UI mockup.

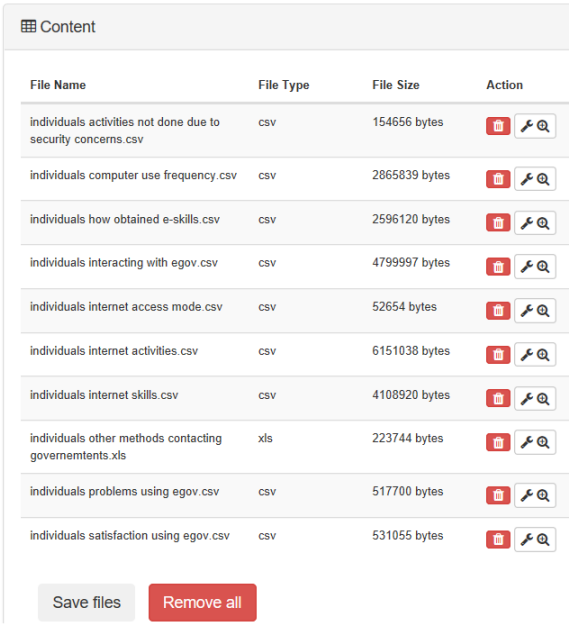

Figure 7: Harvested resources UI mockup.

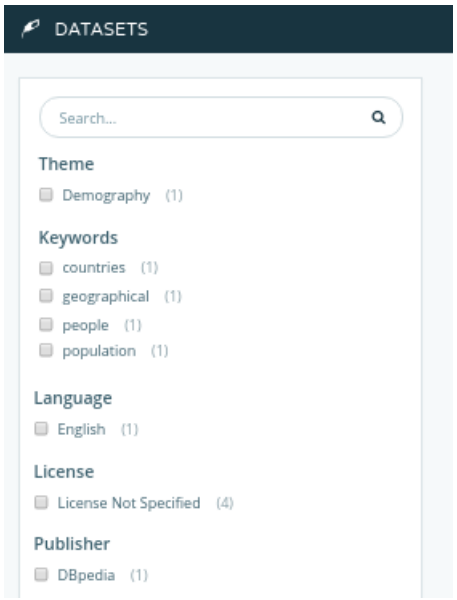

Figure 8: Findability UI example from the DataTank.

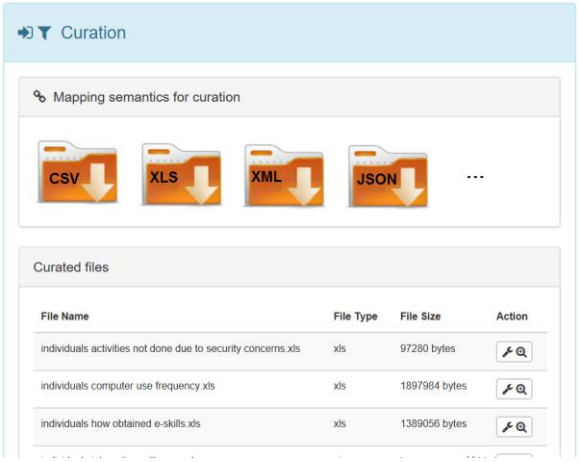

Figure 9: Default curation UI mockup with predefined formats using automated curation (e.g. based on header similarity).

Figure 8 shows an example a search feature of the DataTank that will be used in the DHCF initial version. Figure 9 shows the default curation interface which will allow adapting loaded resources formats, e.g. converting the selected resources into csv, xls, json, xml, … formats.

The DHCF component will have three fusion options. By default if no semantics is defined the resources will be fused based on header similarity. The second option will allow fusing files using a mapping configuration file and/or RML mapping definitions (Figure 4, 5).

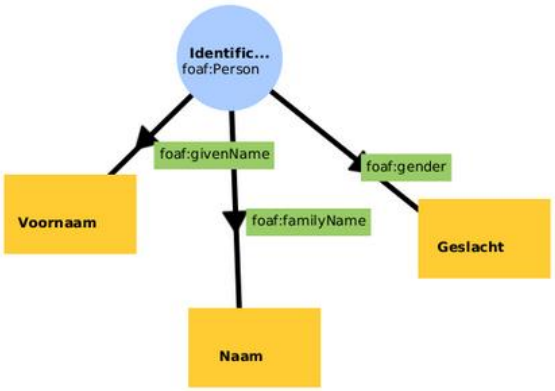

Figure 10: Example of a graphical interface for mapping with RML editor (Dimou et al., 2014).

A more advanced version will allow graphical interface to facilitate the mapping process using RMLMapper. An example of defining semantic mapping is showing in Figure 10.

The unified dataset will ultimately serve an input to KPI visualization engine to support recommendations for the improvements of services in general or allowing zooming into a specific service/user context.

4 VALIDATION MECHANISMS

With respect to the mapping methodologies the proposed method will benefit from the quality assurance mechanism based on semantic linked data technologies. In the case of the tool RML, the quality assessment process that will be followed is the one done for Linked Data so as data owners do not need to maintain and learn multiple tools. RML has achieved that by extending RDFUnit (AKSW, «RDFUnit») which is (one of) the pioneer tools for this job. RDFUnit (and consequently the mapping rules that generate the Linked Data assessment approach) mainly focuses on semantic annotations quality assessment rather than on data values.

With respect to data management the DHCF component in the context of CITADEL ecosystem will follow the principles of FAIR (FAIR Data Management, 2016) data to enable open access, searchability, interoperability and re-usability of the data resources.

With respect to evaluation of the architecture for recommendation purposes an empirical approach will be followed to test the approach in the context of several use cases using national, regional and local e-Government portals described in CITADEL project. The implemented solution will offer built-in evaluation mechanisms considering constructs from commonly accepted technology acceptance models that would also allow to keep track of user perceptions and preferences.

5 CONCLUSIONS

In this work we presented an approach to support recommendations for the utility and improvement needs for public service. We achieve this by facilitating the process of collection, curation and fusion of data originating from various sources in different formats that may provide broader access to KPIs relevant for public service improvement needs. Thanks to the use of linked data technologies the semantics between different sources and formats can be maintained. The fused output will be ready to be processed by analytics and visualization engines to produce suggestions at different levels (e.g. national, regional, local).

The approach also demonstrates a potential for the use of personalized recommendations based on individual profiling, for instance collecting user variables from various sources and matching with existing service catalogues, as well as suggesting issues, improvements needs in the procedures, as well as opportunities for new services, e.g. based on collected data on user feedback, social media activities, information on offline visits/inquiries.

Among the potential limitations of the work can be listed the fact that the architecture does not provide immediate mechanism for addressing privacy, ethical and legal aspects related to data collected from sources other than open public data repositories, which constitutes further research area. In the more advanced version of the model this will be covered by the privacy and security components of the CITADEL ecosystem that will among others deal with anonymization, data encryption/ decryption mechanisms. Yet another concern that suggests further research is potential conflicts between different licence policies that may arise from the use of datasets originating from various sources. While in principle, CITADEL fosters CC0 open data licensing scheme, use of third party data may need further approaches to be researched. In addition this work does not report on the specifics of the implementation of the proposed architecture such as scalability and performance aspects, capture of changes in data over time, etc. While the components and related technologies used in the proposed architecture in principle enable these dimensions, these specific topics remain beyond the

scope of this paper. These will however be covered in the extended version of the work.

Another direction for future work, as already mentioned above, includes empirical studies for evaluations of the usability aspects of the component as well as the impact of recommendations that can be achieved by exploiting the proposed architecture in the context of a recommender system.

The more mature version of the design proposed in this work should also consider built-in mechanisms for capturing end-user perceptions as user acceptance can be important to ensure the effectiveness and continuous refinements needs and ultimately determine its intended utility.

Furthermore, not many studies can be found in the domain of feedback automation (Sedrakyan 2016; Sedrakyan, Snoeck, 2016). Thus methodologies and frameworks to extend recommendations beyond visualization techniques by the use of automated textual feedback targeting both facilitation of interpretability of data visualizations as well as procedural suggestions (Sedrakyan, Snoeck, 2017) will constitute further research direction.

Although the focus of CITADEL project is limited to PAs and public services, the approach can be also inspirational beyond the domain of e-government for the generic context of recommender systems.

ACKNOWLEDGEMENTS

This work has been supported by EC funds from CITADEL project - Empowering Citizens To Transform European Public Administrations (H2020-SC6-CULT-COOP-2016-2017, EC Grant Agreement 726755).

REFERENCES

AKSW, «RDFUnit», Available: http://aksw.org/Projects/RDFUnit.html. Accessed 27.11.2017.

Al-Hassan, M., Lu, H., Lu, J., 2015. A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system. *Decision Support Systems*, *72*, pp.97-109.

CITADEL Consortium, 2017. CITADEL Project. [Online] [Accessed 04 October 2017].

CITADEL Consortium, 2017. *D4.3 Initial CITADEL Ecosystem Architecture*, s.l.: s.n.

Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E. and Van de Walle, R., 2014, April. RML: A Generic Language for Integrated RDF Mappings of Heterogeneous Data. In *LDOW*.

FAIR, European Commission Guidelines on Data Management in Horizon 2020, Version 3.0, 2016. Accessed 27.11.2017.

Heitmann, B., Hayes, C., 2010. Using Linked Data to Build Open, Collaborative Recommender Systems. In *AAAI spring symposium: linked data meets artificial intelligence* (pp. 76-81).

Sedrakyan, G., 2016. Process-oriented feedback perspectives based on feedback-enabled simulation and learning process data analytics. PhD Thesis, KU Leuven.

Sedrakyan, G., Snoeck, M., 2016. Enriching Model Execution with Feedback to Support Testing of Semantic Conformance between Models and Requirements-Design and Evaluation of Feedback Automation Architecture. In *AMARETTO@ MODELSWARD* (pp. 14-22).

Sedrakyan, G., Snoeck, M., 2017. Cognitive feedback and behavioral feedforward automation perspectives for modeling and validation in a learning context. In *Model-Driven Engineering and Software Development* (pp. 70-92). Springer, Cham.

Vander Sande, M., Colpaert, P., Van Deursen, D., Mannens, E. and Van de Walle, R., 2012. The DataTank: an open data adapter with semantic output. In *21st international conference on world wide web*, proceedings (p. 4).

W3C, 2014. RDF 1.1 concepts and abstract syntax.

W3C, 2013. SPARQL query language for RDF. W3C Recommendation.