**Advanced Usage of Integrated Data in Public Sector**

**Vuckovic Milos**

**mvuckovic70@gmail.com**

**15.04.2016**

# Abstract

Public sector, as a major carrier and collector of information, faces numerous and complex challenges: proliferated global security risks, social inequities and inequalities, traffic incidents, suicides, social inclusion problems, inadequate care of children with special needs and elderly people, etc.

Great deal of such unwanted outcomes that could have been predicted and prevented is attributed to human activities manifested in late or inadequate response. Hence, we could also state that various challenges in a public sphere occur as the aftermath of bad interoperability and lack of data connectivity and integration between government structures.

Let us suppose, however, that we can create the mechanism of integrating behavioral, psychological, educational, health, work, criminal, forensic, social media and other records into one data model, both robust and flexible, individually adjusted, yet systematized and categorized following certain rules, connected across horizontal and vertical public domain structures, well documented and referenced, with major objectives: guide, simplify, discover, predict, prevent.

Such a system, supported by various powerful statistical analytical tools, and strengthened by contemporary scientific models, would likely cut public expenses, enhance security and reduce the probability of unwanted outcomes to occur.

The main goal of this paper is twofold: to support preemptive and proactive activities of public managers and to provide extended assistance to its users that go beyond traditional online services using advanced public data integration framework.

Another idea is to leave data open to a certain point, which does not automatically exclude creation of open data portals with information available to a wide range of users (e.g. to enable companies to develop their applications based on such data).

**Keywords**: Data integration, information system, interoperability, collaboration, public sector, predictive analytics, actionable intelligence

# Table of contents

[Abstract 1](#_Toc450817808)

[Table of contents 2](#_Toc450817809)

[1. Introduction 3](#_Toc450817810)

[2. Literature review 4](#_Toc450817811)

[2.1. Public data integration - terms and prerequisites 4](#_Toc450817812)

[2.1.1. E-Government concept 4](#_Toc450817813)

[2.1.2. Interoperability and cross-sector collaboration 4](#_Toc450817814)

[2.1.3. Integrated Data System and Actionable Intelligence 5](#_Toc450817815)

[2.1.4. Data mining 5](#_Toc450817816)

[2.2. Government organization 6](#_Toc450817817)

[2.3. Major issues in public data integration concept 6](#_Toc450817818)

[2.3.1. Functional and organizational issues 7](#_Toc450817819)

[2.3.2. Legal, privacy, security and ethical issues 7](#_Toc450817820)

[2.3.3. Technology and methodology issues 8](#_Toc450817821)

[3. Research Objectives and Approach 9](#_Toc450817822)

[3.1. Research goals 9](#_Toc450817823)

[3.2. Research methodology 9](#_Toc450817824)

[3.2.1. Background 9](#_Toc450817825)

[3.2.2. Steps to take 10](#_Toc450817826)

[3.2.3. Determining inter-sector correlations 10](#_Toc450817827)

[3.3. Additional approach 10](#_Toc450817828)

[4. Current Work and Preliminary Results 11](#_Toc450817829)

[4.1. Current work 11](#_Toc450817830)

[4.1.1. Defining project goals 11](#_Toc450817831)

[4.1.2. Defining sectors of interests 11](#_Toc450817832)

[4.1.3. Determining dimensions correlations and distributions 11](#_Toc450817833)

[4.1.4. Choose data integration technique and tools 11](#_Toc450817834)

[4.1.5. Determining data sources 12](#_Toc450817835)

[4.1.6. Defining and conducting rules of transformation 13](#_Toc450817836)

[4.1.7. Ranking methodology 15](#_Toc450817837)

[4.1.8. Visualizing results 16](#_Toc450817838)

[4.2. Preliminary results 18](#_Toc450817839)

[5. Work Plan and Implications 20](#_Toc450817840)

[6. Conclusion 21](#_Toc450817841)

[References 22](#_Toc450817842)

# 1. Introduction

The very existence of integrated data concept in public domain is good enough for citizens to pay their taxes online, get new passport without visiting the public office or check the school grades of their kids electronically. The true challenge hides in question: can we do more with the same data on individual or entity level?

Such system should be capable of personalization, matching citizens’ and businesses’ circumstances and needs. Citizen-centric design is dependent on a fully integrated operational model usually requiring significant systems integration and accompanying transformation of business processes (UN, 2014).

While there were many individual ideas on how to improve cross-sector collaboration, such as in elderly care domain (Grundischi, 2013), or using various solutions of big data usage in many real-life situations (Hopfgartner, 2015), this paper would try to raise the question over advanced usage and transformation of existing public data into knowledge by fostering the creation of the mechanism with embedded capability of providing global trends or forecasts on various levels (local, state, industry-specific, group-specific, etc.).

The primary goal of this project is to examine various practical examples of actionable intelligence by using different scientific methods for uncovering hidden relations within data, showing the ways to deliver benefits, both practical and financial, to individuals and to public sector in general.

The additional objective is to expand the integrated data system usage to business users, as well as to government as a whole.

The usage of integrated data system concept faces numerous challenges and issues of different types to tackle – legal (in terms of standards to follow), ethical and security (to ensure that research field based on personal data is risk free), technological (data integration, data quality, transformation, matching and linking procedures), financial (with objective to minimize costs of data acquisition, integration and manipulation), functional and so on.

Current major issues will be debated more thoroughly in chapter 2, as well as the brief literature review, followed by the detailed analysis of main government cross-sector collaboration and interoperability tasks and issues.

This paper will then address the terms of actionable intelligence and integrated data systems and the ways they grasp the added value of such data.

# 2. Literature review

## 2.1. Public data integration - terms and prerequisites

### 2.1.1. E-Government concept

The idea of public data integration originally derives from the concept of E-government.

There are many different definitions of E-government. According to UN, “E-government is defined as utilizing the internet and the world-wide-web for delivering government information and services to citizens.” E-government implementation and public information integration is generally looked as a driving force to promote the one-stop service transformation” (Zhang et al., 2011).

Evolutionarily speaking, E-government is going through its fifth stage called the networked presence, preceded by emerging presence (limited information), enhanced presence (more direct and selective information), interactive presence (more complex forms of unidirectional interactions) and transactional presence (two-way interaction), respectively.

Current stage of e-government development is reflected in integration of all transactional public services through the so called “whole-of-government approach” (UN, 2012).

### 2.1.2. Interoperability and cross-sector collaboration

Conditio sine qua non of integrated concept of e-government is existence of cross-sector collaboration and data interoperability. “To enhance the efficiency and effectiveness of public service delivery and create more public values for the citizens through making the best use of the potential of ICT, cross-sector integration and collaboration is becoming an imperative for the public managers” (Nan, 2012).

Any number of complex public problems would appear to require cross-sector collaboration focusing on multiple levels of decision making and action from top-level policy-making bodies to individuals (Bryson et al., 2006)

Collaboration is a process in which autonomous actors interact through formal and informal negotiation, jointly creating rules and structures governing their relationships and ways to act or decide on the issues that brought them together; it is a process involving shared norms and mutually beneficial interactions (Thomson & Perry, 2006). Cross-sector collaboration is the linking or sharing of information, resources, activities, and capabilities by organizations in two or more sectors to achieve jointly an outcome that could not be achieved by organizations in one sector separately. (Bryson et al, 2006). Interoperability, as a direct and most positive consequence of collaboration, is “the ability of government organizations to share and integrate information by using common standards” (Gottschalk, 2009).

### 2.1.3. Integrated Data System and Actionable Intelligence

With advent of public data integration concept, a prominent line of research now focuses on a term called actionable intelligence. Actionable intelligence (AI) is derived from relevant contributors working together to make meaning out of data produced, formulating a theory of change, and taking action to test this formulation (Fantuzzo, 2013). The heart of AI is Integrated Data System (IDS), which integrates individual citizens’ data across agencies to provide scientifically sound, intergovernmental information (Fantuzzo, 2015).

IDS typically contains data collected as a part of the day-to-day business of operating government public assistance, social service, health, and/or education programs. They link individual data across multiple, independent agency data systems allowing service access and use patterns to be tracked across service delivery systems (Limlingan, 2015). Data in IDS is integrated and then stratified according to specific objectives of each sector.

IDS is a huge paradigm shift towards discovery of new ways of integrated government data usage. As a relatively new concept it should provide an in-depth analysis of cross-sector integrated public data and enable creation of powerful tools for systematic approach. Although being in an embryonic stage, “there are some organizations that have already set up functioning integrated data systems capable of producing actionable intelligence to guide policy. These include the states of Michigan and South Carolina, and the County of Los Angeles” (Culhane, 2016).

Several organizations throughout the US have independently developed their own integrated data systems. These are projects led by state governments, local governments, and universities (Culhane, 2010). All those projects are using integration data pattern for particular project within a specific sector, e.g. early childhood IDS (Cochenour, 2015), or IDS for Low-Income Hispanic Families with Young Children (Limlingan et al, 2015).

### 2.1.4. Data mining

From government census to personnel and customer files, very large collections of information are continuously gathered about individuals and groups. Governments, companies and organizations such as hospitals, are stockpiling very important quantities of personal data to help them manage human resources, better understand a market, or simply assist clientele. Regardless of the privacy issues this type of data often reveals, this information is collected, used and even shared. When correlated with other data, this information can shed light on customer behavior and the like (Zaïane, 1999).

Translating the exponential growth of new data into actionable decision-making requires high-value analytical skills which exist only in few governments. Therefore, significant attention must be placed on data mining, (Michael E. Milakovich et al., 2013). Fine-grained predictive models on datasets will leverage the true potential of such approach. Following example clearly demonstrates impact of data integration: In terms of cost savings in health-care sector in USA thanks to big data practices and integration methodologies implemented, “the pathways could account for $300 billion to $450 billion in reduced health-care spending, or 12 to 17 percent of the $2.6 trillion baseline in US health-care cost.” (Groves et al., 2013).

Data mining concepts are already being widely explored in different forms for various public domains: in national security (Zanasi, 2008 and Thuraisingham, 2004); for tax and customs audit (Cleary, 2012), for geospatial analysis (Hollywood et al., 2011); in education (Merceron, 2006 and Baradwaj, 2011), etc.

To conclude, this continuous cross-sector connectivity chain imposes data integration as a top factor in making important decisions, which delivers benefits both practical and financial, by saving the time of decision making process, cutting the strategy implementation costs, and in the same time providing useful information suitable for further use and deeper analysis.

## 2.2. Government organization

We are living in an era of massive data explosion, which does not show the signs of slowing down. Instead, it is speeding up in all directions. One of the potentially blossoming areas of such development is the public sector.

Government apparatus is probably the most complex human-made system in terms of amount of information being obtained and processed. Public sectors operations under governmental umbrella are spread across vast aspects of human existence, activities, environment and related affairs, or events. It is almost impossible to comprehend the magnitude and variety of data flowing towards and from the public administration offices and, consequently, even more difficult to cope with such complexity.

Governments should foster the cross-sector collaboration and data exchange with aim to increase data interoperability in general. Used in a proper manner, available integrated data could be transformed into knowledge by uncovering hidden relations, rules and behaviors, thus adding to decision making process in public administration.

In a world of ubiquitous connectivity, the governments cannot afford to lose pace with global trends. Instead, they should boost the cross-sector interoperability and induce new collaborative concept of new, integrated public information system framework.

However, task of integrating and making sense out of such data is far from simple and it faces challenges that often go beyond existing technology, hindered by the lack of common framework for different data types and sources. Mastering the creation of such environment is the paramount and sufficiently challenging assignment with crucial questions being posed: is it possible to integrate such amount of data in a meaningful way? What data actually *could* mean?

## 2.3. Major issues in public data integration concept

Data integration is a process carrying loads of challenges. Applied in a governmental environment, the burden gets even heavier.

Public sector is complex enough to dig deep in integration procedures and mechanisms. It is entangled resource- and time-consuming process, requiring full commitment on collaborative basis. The explosion of compound big data corpus and internet of things adds to the complexity even more.

Current governmental data integration achievements mostly pertain to data integration within an individual public sector. In many cases each public sector has the potential to anchor dozens of informational systems of different types, structures, data sources, transactions, reports.

There has been also a raising concern over passive role of the government agencies with critics of being reactive (react to an outcome) instead of being proactive (prevent the outcome) in many areas of their interest.

There are a number of political, organizational and technical challenges, which may hinder a more collaborative approach in government and with outside actors (UN, 2014):

* Lack of a coherent vision and commitment to address sustainable development issues;
* Weak collaborative leadership and “silos” like mentality;
* Entrenched power structures;
* Vertical and horizontal organizational fragmentation;
* Inadequate accountability mechanisms for cross-agency collaboration and where appropriate, minimum or “appropriate” quality standards;
* Mistrust among ministries/agencies;
* Lack of confidence in the IT infrastructure, data privacy and security.

The comprehensive analysis of risks in integrated data system concept has been brought by David Culhane et al. in a work “Connecting the Dots”, 2010. Most of the abovementioned work is summarized in this paper as well.

### 2.3.1. Functional and organizational issues

The functional problem could be defined as functional incoherence of public sectors and agencies, inert and insufficiently educated administration managers. Four major dysfunctions of public problem solving and decision making are identified within US government structures that thwart effective, efficient, and ethical public services: (1) top-down, one-way, hierarchical leadership; (2) compartmentalized bureaucracies with rigid boundaries; (3) undisciplined decision making that is purely reactive and politically motivated; and (4) disconnects between “knowing” and “doing” communities. Despite the increasing interest in cross-sector collaboration, these partnerships were plagued with two major problems: (1) inadequate managerial structures and (2) inadequate managerial process. (Fantuzzo, 2015).

### 2.3.2. Legal, privacy, security and ethical issues

Most of ethical and security issues refer to data privacy and confidentiality within healthcare informational systems, social services (youth protection, childcare data, people with disabilities, elderly people care, housing problems, etc.), but also to security of every individual public data exposed to risk of being compromised or misused.

A number of complex legal issues must be considered relating to the privacy of persons within these systems. The rights to use various types of data for research are regulated by federal law, state law, and public access policies (Culhane, 2010).

In the United States, there are no distinct guidelines for states on how to implement this privacy protection, and each individual state must demonstrate to the federal government how it provides privacy protection (Culhane, 2010).

Standards for protecting the privacy of individually identifiable health information were established by the United States Department of Health and Human Services (HHS). Research with human subjects should be conducted in the best interest of the individual participants, safeguarded by their informed consent or the permission of officials charged with ensuring the confidentiality of their administrative records (Culhane, 2010).

### 2.3.3. Technology and methodology issues

The ultimate usefulness of administrative data for research and planning purposes depends on the original data quality. Systems administrators must develop data acquisition, auditing, and linkage procedures that assure data’s integrity for research purposes. This is the science of integrated data systems (Culhane, 2010).

In terms of data integration technology and methodology, it proved as extremely difficult to integrate public data into one meaningful system, due to many factors: diversity of information systems across different public sectors, lack of integrating mechanisms, the rise of big data and its complexity, vertical and horizontal fragmentation, and many others.

The methodology of integrating data is not unified. Some approaches rely on schema matching, some on creating individual database triggers to push the selective data to an ETL process and some are using traditional data warehouse creation procedures.

Traditional silo-concept of data manipulation is an omnipresent issue difficult to cope and transform. Although no country’s portal completely integrated all information, services, and features assessed, several came close. Some of these vanguard countries include: the Republic of Korea, the United Arab Emirates, and the United Kingdom (UN, 2012).

Computer scientists have developed methods for addressing many of these issues from a technical standpoint. These methods can range from complex real-time (or nearly real-time) relational databases with sophisticated record linkage systems (for finding unique identifier for each person), to more basic parallel archival processes with annual updates that are merged using stored record linkage procedures (Culhane et al, 2010).

Records linkage may occur as another type of problems in data architect’s endeavors of finding unique identifier on individual level. ”IDS implementations typically acquire data from more than two data sources. Some may have as many as ten or more data sources. This presents the problem of handling a “link cascade” where record A and record C might be linked to each other not because they match each other, but because they each match record B.” (Kumar, 2015).

# 3. Research Objectives and Approach

## 3.1. Research goals

The overarching objectives of this paper are following:

* To use Integrated Data System to create Actionable Intelligence environment
* To point at additional benefits of using existing public records in general, which should ultimately show new outcomes of cross-sector data integration;
* To cut public exploration costs and need for expensive additional researches;
* To pinpoint the directions public managers should act towards, thus reaping societal benefit from these actions
* To widen methodology approach by using data classification method which is mostly being used in business sector up to now
* To expand the same method within public sector referring to business and G2G usage

This approach will go beyond data integration on individual level. Moreover, it will use anonymized (to avoid ethical and privacy issues) individual cross-sector records to create a global picture of a certain process by grouping them following certain rules and principles.

Although perceived as a very complex process, it could be simplified without losing its analytical power, by targeting only records and tables that directly affect the desired result.

## 3.2. Research methodology

### 3.2.1. Background

This paper relies on Customer Stratification methodology developed by a group of scientists within Texas A & M University and deployed in wholesale and distribution business to optimize many business-related activities. It also became a part of SAP Hana Business Intelligence package. This method classifies customers within a business entity using four-dimensional classification based on (a) sales volume, (b) gross margin, (c) cost to serve, and (d) loyalty (Lawrence et al., 2011). It joins them into a stratification quadrant and then separates the customers into 4 groups using results obtained.

This method uses various business transaction tables (sales, CRM, inventory, costs), puts them together and conducts various statistical calculations in order to get score-like results for each pre-defined parameter or for each customer group, respectively.

The results are visualized in a scatter matrix and showing several aggregated trends during some period of time, and pointing the individual customer position and its transactional data within it in the same time, with clear picture what to do with it. The main purpose of this approach is to define the business strategy for different levels: for particular customer(s), sales managers, at location level, business unit or sales sector as a whole.

### 3.2.2. Steps to take

This work will try to explore the ways of implementing mentioned approach in public sector as well. Research goals mentioned in paragraph 3.1 require conducting the following steps:

* Identifying goal(s) of the project
* Identifying sector(s) of interests
* Identifying correlation(s) between chosen sectors
* Identifying correlation distribution between sectors
* Identifying targeted transactional source(s) of each sector
* Identifying table(s) of interest per source
* Identify distribution frequency by tables
* Implementation of cross-sector data integration process
* Periodical update setting in months
* Conducting a classification method to get results for:
  + Aggregation
  + Visualization
  + Presentation
* Implementation of various strategies to various data groups
* Monitoring

Each of these steps will be examined and processed more detailed in section 4.

### 3.2.3. Determining inter-sector correlations

The main overall challenge of this proposed approach is to find matching public areas for classification and determine their correlation, in order to get accurate and actionable results at the end of the whole process.

Data from four sectors will be joined in order to find patterns of interconnectedness which would ultimately lead to a scientific model: health care, education, social and criminal records. Several independent studies have shown a positive correlation between some of them. “Research has uncovered a large and positive correlation between education and health” (Lleras-Muney, 2004). “Crime rates and inequality are positively correlated and it appears that this correlation reflects causation from social inequality to crime rates, even controlling for other crime determinants” (Fajnzylber, 2002).

## 3.3. Additional approach

In terms of approach implementation viability, the same pattern could likely be used for public data relating to companies containing main balance positions, profit, loss, number of employees, payables and receivables. The same principle also counts for public sector divisions as well, its employees and activities.

# 4. Current Work and Preliminary Results

## 4.1. Current work

Current work is based on steps previously mentioned in paragraph 3.2, which will be processed more detailed in sub-paragraphs to follow.

### 4.1.1. Defining project goals

Primary goal of the project can be driven by some research or public initiative which expresses the need for advanced use of current public data, for example: to create various government policies on different levels and groups using cross-sector integrated data system.

### 4.1.2. Defining sectors of interests

Sectors of interests are as follows: healthcare, education, social sector, criminal records. As of now, for the exploration purposes, sectors will be named dimensions.

### 4.1.3. Determining dimensions correlations and distributions

Following correlation pattern is presumed:

* Positive correlation between education and healthcare exists (matrix EH)
* Positive correlation between social status and criminal records exists (matrix SC)
* Positive correlation exists between two correlation matrices

Correlation distribution between dimensions represents percentage participation of each dimension in final matrix rank. In this paper it is presumed as follows:

Table - Relative importance within correlation matrices

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 2 | 3 | 4 |
| **HEALTHCARE** | **EDUCATION** | **SOCIAL STATUS** | **POLICE RECORDS** |
| MATRIX EH | | MATRIX SC | |
| 50% | 50% | 30% | 70% |

### 4.1.4. Choosing data integration technique and tools

Data integration process and methodology will be the subject of particular interest for our future research steps. Data integration is based on IDS concept with main objective to create actionable intelligence area. Different data integration technologies and tools are being used to conduct creation of new analytic environment. Modern data integration practice identifies three types of data integration (Kumar, 2015.):

* Need Based Data Integration (project-specific)
* Periodic Data Integration (for a class of business problems or analytic needs
* Continuous Data Integration (address certain complexities that result from the real-time processes)

Primary data integration steps are as follows:

* Data profiling (data quality analysis)
* Data cleansing (format standardization, assigning missing parts of a record)
* Identity matching (using social security number, ID card, name, year of birth, etc.)
* ETL procedures (data extraction from the source, data transformation, aggregation and loading into new analytical environment)
* Upserts (update if any change in existing records, or insert if new data occurs in a source system), by pull approach (using ETL) or push approach (using source DBMS triggers or procedures)

Data integration process should involve following database entities:

* Source database(s)
* Stage database (simple loading of all source tables into a new, temporary database ready for further transformation)
* NDS - normalized data store as a result of ETL procedures on staging database
* DDS - dimension data store, de-normalized database, ready for further analysis

### 4.1.5. Determining data sources

This part of the process involves identification of source transactional tables for each individual and each originating sector respectively. All these tables and records will be extracted from source sectors into the stage database following particular data linking rules. Data linking represents a process where individual records are being matched across disparate administrative data systems (Kumar, 2015). The goal of data linking is to create unique client identifier for all sectors and, consequently, new integrated data entity.

Table – Source transactional tables per dimensions

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **HEALTHCARE** | | **EDUCATION** | | **SOCIAL** | | **POLICE RECORDS** | |
| MEDICAL ASSESSMENT | | EDUCATION LEVEL |  | ASSETS |  | ENVIRONMENT RANK | |
| RISK GROUP | | INDUSTRY |  | LIABILITIES |  | CRIMINAL RECORDS | |
| HEALTH CATEGORY | | INDUSTRY CATEGORY | | SPENDINGS | | CRIMINAL CATEGORY | |
| PSYCHO TEST RANK | | AVG GRADES |  | SAVINGS |  | OFFENCES NUMBER | |
| IQ TEST RANK | | BEHAVIOR RANK | | HOUSING |  |  |  |
|  |  | SCHOOL RANK | | COMUNAL COSTS | |  |  |
|  | |  | | WORK EXPERIENCE | |  |  |
|  | |  | | JOBLESS PERIOD | |  |  |
|  | |  | | INCOME LEVEL | |  |  |
|  |  |  |  | LOCATION |  |  |  |

Process of determining source table starts with register containing basic individual data:

Table - Birth and residency register

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **birth register** | data type | foreign table | sample value |  |
| **ID** | **identity, int** |  | **654213484** |  |
| firstName | varchar |  | Robert |  |
| lastName | varchar |  | Schweizer |  |
| genderID | integer |  | M |  |
| Birthdate | date |  | 07.12.1965 |  |
| parent1ID | integer | birthReg | Michael |  |
| parent2ID | integer | birthReg | Eva |  |
| birthPlaceID | integer | locationReg | Munchen, GER |  |
| residencyID | integer | locationReg | Munchen, GER |  |
| residentStatusID | date | residencyReg | Permanent |  |
| familyStatusID | integer |  | Father |  |
| countryBirthID | integer | countryReg | Germany |  |

Source tables from chosen sectors might look as follows:

Table – Source tables of each sector

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **healthcare register** |  | **education register** |  | **social register** |  | **criminal register** |
| ID |  | ID |  | ID |  | ID |
| age |  | degreeID |  | employed |  | heavyCrimesRatio |
| weight |  | titleID |  | salaryMonthlyNett |  | mediumCrimesRatio |
| height |  | grade |  | salaryAVG |  | lightOffencesRatio |
| bloodTypeID |  | institutionTypeID |  | workExperience |  |  |
| bloodQuality\_HGB |  | institutionRank |  | joblessPeriod |  |  |
| urineQualityCat |  | institutionID |  | liabilities |  |  |
| chestRayCat |  |  |  | assets |  |  |
| ekgCat |  |  |  |  |  |  |
| IQ\_Cat |  |  |  |  |  |  |
| psychoTestCat |  |  |  |  |  |  |
| disabilityCat |  |  |  |  |  |  |
| previousTreatmentsCount |  |  |  |  |  |  |
| diagnosysID |  |  |  |  |  |  |

### 4.1.6. Defining and conducting rules of transformation

It will be necessary to perform transformations during the integration process following defined rules, because source values should be transformed into ranks for each parameter within each table, while preserving initial data for individual or grouped end-to-end analysis:

1. Identification of particular entity (table) distribution. Distribution of importance serves as a mechanism for a final dimension ranking. Table below represents estimated importance of each extracted table within a healthcare dimension:

Table – Determining relative importance within dimension entity

|  |  |
| --- | --- |
| **HEALTHCARE** | **RELATIVE IMPORTANCE WITHIN SAME DIMENSION** |
| PSYCHO AND IQ TESTS RANK | 10% |
| MEDICAL ASSESSMENT | 40% |
| RISK GROUP | 30% |
| HEALTH CATEGORY | 20% |

1. Loads will be on periodic basis, since we want to analyze trends from periodical changes as well:

Table – Periodical update of each dimension in months

|  |  |  |  |
| --- | --- | --- | --- |
| **HEALTHCARE** | **EDUCATION** | **SOCIAL** | **POLICE RECORDS** |
| 6 | 12 | 6 | 12 |

1. Rules of transformations will be implemented either during ETL process or by creating new DDS, depending on whether we want to keep some of original values or not. Some columns are representing minimum or maximum values, and some of them are classified within certain ranges in order to get a ranking for each relevant column. Depending on distribution rules, the individual column ranks are used to form an overall rank within a particular dimension:

Table – Samples of range values transformation

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **medical: bloodQuality\_HGB** | | |  | **education: grade** | | |  | **social: salaryAVG** | | |  | **criminal: lightOfeencesRatio** | | |
| **name** | **value** | **rank** |  | **name** | **value** | **rank** |  | **name** | **value** | **rank** |  | **Count** | **value range** | **rank** |
| min | 110 |  |  | min | 1 |  |  | min |  |  |  | > 20 | > 20% | D |
| max | 180 |  |  | max | 10 |  |  | max |  |  |  | 15 – 20 | 15 % - 20 % | C |
| range | 120 - 170 | A |  | range | 1-5 | D |  | range | < 1.200 | D |  | 10 – 15 | 10 % - 15 % | B |
| range | 100 - 120 | B |  | range | 6-7 | C |  | range | 1.200 – 2.000 | C |  | < 10 | 0 %- 10 % | A |
| range | 170 - 190 | B |  | range | 7-8 | B |  | range | 2.000 – 3.500 | B |  |  |  |  |
| range | 90 – 100 / 190 -200 | C |  | range | 9-10 | A |  | range | > 3.500 | A |  |  |  |  |
| range | < 90 or > 200 | D |  |  |  |  |  |  |  |  |  |  |  |  |

1. After the completion of the first phase of transformation, the targeted tables will take following shapes, respectively:

Table – Education register ranking

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **education register** | data type | foreign table | sample value | distribution | rank |
| ID | identity, int |  | 654213484 |  |  |
| degreeID | integer |  | master | 50% | A |
| titleID | integer | titlesReg | engineer | 10% | A |
| Grade | numeric |  | 8.42/10 | 20% | A |
| institutionTypeID | integer | instTypeReg |  | 10% | A |
| institutionRank | integer |  |  | 10% | A |
| institutionID | integer | institutionReg | |  |  |
| **finalRank** | **integer** |  | **0.91** |  | **A** |

Table – Health care register ranking

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **medical register** | data type | foreign table | sample value | distribution | rank |
| ID | identity, int |  | 654213484 |  |  |
| Age | numeric |  | 51 |  |  |
| Weight | numeric |  | 91 |  |  |
| Height | numeric |  | 184 |  |  |
| bloodTypeID | integer |  | A+ |  |  |
| bloodQuality\_HGB | integer |  | 92% | 10% | A |
| urineQualityCat | integer |  | 75% | 10% | B |
| chestRayCat | integer |  | 95% | 10% | A |
| ekgCat | integer |  | 89% | 10% | A |
| IQ\_Cat | integer |  | 126 | 10% | A |
| psychoTestCat | integer |  | 92% | 10% | A |
| disabilityCat | integer |  | 0% | 10% | A |
| previousTreatmentsCount | integer |  | 4 | 20% | B |
| diagnosysID | integer | diagnosysReg | 1.67 | 10% | B |
| **finalRank** | **integer** |  | **0.84** |  | **A** |

Table – Social register ranking

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **social register** | data type | foreign table | sample value | distribution | rank |
| ID | identity, int |  | 654213484 |  |  |
| employed | bit |  | YES | 20% | A |
| salaryMonthlyNett | numeric |  | 1,436.00 | 20% | C |
| salaryAVG | numeric |  | 64% | 20% | C |
| workExperience | numeric |  | 26 | 10% | A |
| joblessPeriod | numeric |  | 1.2 | 10% | A |
| liabilities | numeric |  | 145% | 10% | C |
| assets | numeric |  | 41% | 10% | C |
| **finalRank** | **integer** |  | **0.51** |  | **C** |

Table – Criminal records ranking

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **criminal register** | data type | foreign table | sample value | distribution | rank |
| ID | identity, int |  | 654213484 |  |  |
| heavyCrimesRatio | numeric |  | 0% | 60% | A |
| mediumCrimesRatio | numeric |  | 0% | 30% | A |
| lightOffencesRatio | numeric |  | 9% | 10% | A |
| **finalRank** | **integer** |  | **0.92** |  | **A** |

### 4.1.7. Ranking methodology

In the next integration step, new integrated table will be created containing only final ranks for each person and each sector, respectively, thus actually forming new actionable intelligence entity, which might look like this:

Table – Virtual AI entity sample containing all integrated source tables

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **INTEGRATION LEVEL 1 – INTEGRATED DIMENSIONS** | | | **HEALTHCARE RECORDS** | | | **…** | **…** |
| ID | Name | Social\_ID | CATEGORY | RISK | AGE | … | … |
| 1 | Marc Trudeau | 654654 | A | A | 26 |  |  |
| 2 | Robert Simon | 231321 | B | D | 69 |  |  |
| 3 | Maria Lucatelli | 987978 | A | B | 45 |  |  |
| 4 | Jürgen Dietrich | 981514 | D | C | 54 |  |  |

In order to preserve privacy of each individual, the table for final analysis will be anonymized and ranked, while adding parameters which would be useful for deeper analysis (by sector, group, industry, educational level, various statuses, etc.):

Table – Joined dimension ranks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LEVEL 1 – INTEGRATED RECORDS | HEALTHCARE | EDUCATION | SOCIAL | POLICE |
| **ID\_654213484** | A | A | C | A |
| **ID\_987465146** | C | B | A | D |
| **ID\_567908469** | B | A | C | A |
| **ID\_235872355** | B | D | C | A |

Table – Anonymized dimensions ranks with additional parameters

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| LEVEL 2 | **REGION\_ID** | **STATUS\_ID** | **INDUSTRY\_ID** | HEALTHCARE | EDUCATION | SOCIAL | POLICE |
| ID1 | 2 | 1 | 12 | A | B | A | B |
| ID2 | 2 | 3 | 16 | C | B | A | D |
| ID3 | 3 | 2 | 7 | B | A | C | A |
| ID4 | 1 | 1 | 23 | B | D | C | A |

The cumulative ranking will be conducted using correlation matrices distribution rules, as defined in a paragraph 4.1.3, rounding to the nearest rank if necessary:

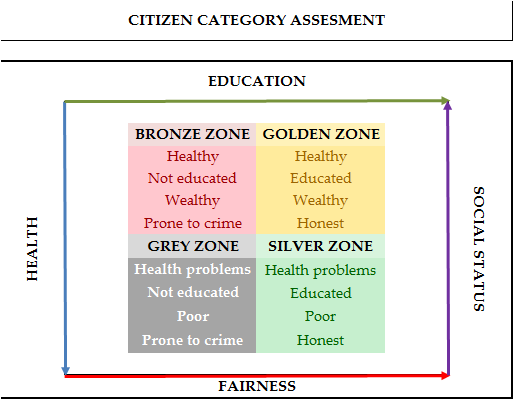
Table – Level 3 – Rules for final rank transformation

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ID | EDUCATION | HEALTH | SOCIAL | FAIRNESS |  | E-H | S-F |  | RANK |
| 1 | D | A | A | B |  | DA | AB |  | CB |
| 2 | B | A | C | A |  | BA | CA |  | AB |
| 3 | A | C | C | A |  | AC | CA |  | BB |
| 4 | D | C | C | B |  | DC | CB |  | CC |
| 5 | D | A | B | B |  | DA | BB |  | CB |
| **6** | **A** | **A** | **C** | **A** |  | **AA** | **CA** |  | **AB** |

### 4.1.8. Visualizing results

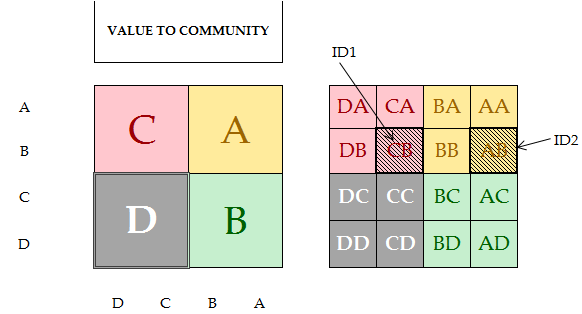
Following step visualizes achieved ranking results using classification quadrant:

Figure - Combining ranking to get record category



This quadrant brings four major zones surrounded by four dimensions carrying main properties of each dimension. Next we bring ranks positions into the quadrant. Each rank is a combination of matrices [healthcare - education], and [social status - criminal records]. Following figure represents ranked distribution in the quadrant with some individual values:

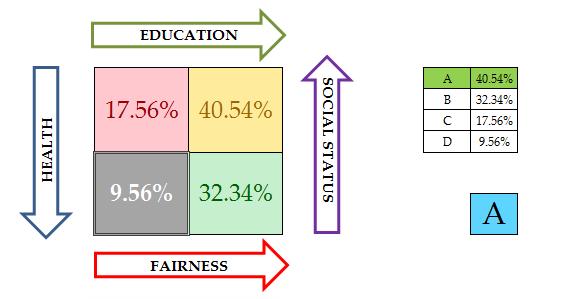
Figure - Finding individual records within the quadrant



## 4.2. Preliminary results

After records being transformed and new entities created and loaded with data and then aggregated, the cumulative classification quadrant might show results like these:

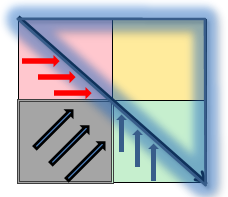
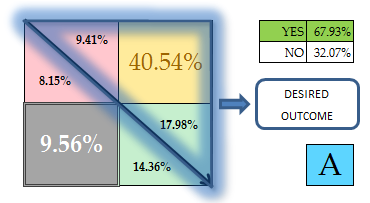
Figure – Quadrant aggregation



The above figure could serve as a good base for analysis and targeted policy. In the figure 4, the marked blue triangle presents the area where we want to bring as many people as we can: well educated, without bigger crime offences, of good health and social status. Since results can be aggregated and grouped by various parameters, public policy which targets only certain groups or regions can be implemented accordingly.

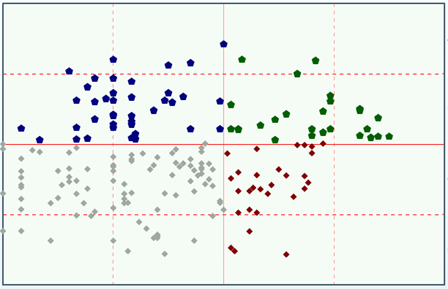
Many preventive and proactive policies can be defined and conducted on various governmental levels as well. In addition, these figures can serve as a base for future activities, or for trend analysis, since records will be periodically updated.

Figure 4 - Public policy recommended areas / policy guidance and directions



Since individual records data are kept for each sector, new system will be ready for individual analysis, interpretation and implementation, or available to a particular individual at request:

Figure – Scatter matrix of all or grouped individual records



This model can develop additional graphical views with more detailed and structural approach and analysis:

Table – Aggregated results distribution by zones and dimensions

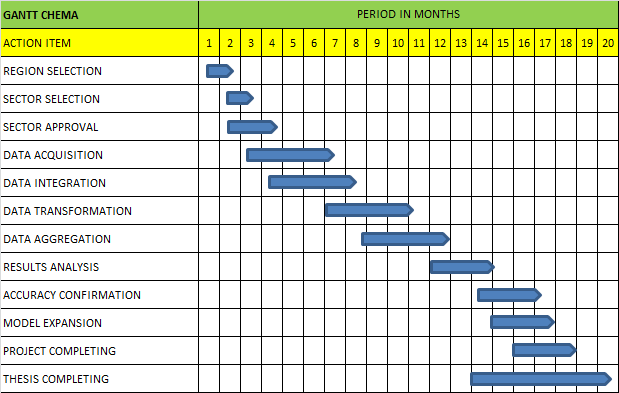
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | GOLDEN | SILVER | BRONZE | GREY |
| HEALTHCARE | 52% | 25% | 16% | 7% |
| EDUCATION | 61% | 19% | 14% | 6% |
| SOCIAL | 59% | 22% | 11% | 8% |
| CRIMINAL | 5% | 13% | 26% | 56% |

Figure – Characteristics of zones by dimension aggregations

# 5. Work Plan and Implications

In this paper theoretical ranking model of cross-sector integrated data is developed. Future research will be phase-oriented and focused on research using the real government data. Given previously mentioned data integration issues, during this process many problems may occur. Nevertheless, it is mandatory to establish specific milestones and timelines of the project development process using Gantt diagram:

Figure – Work plan



This plan does not take into account publication plan, which might include submissions to the doctoral consortium of conferences, publication in journals, submission to international conferences, etc.

# 6. Conclusion

Real value of integrated systems in public sector may stem from using new approaches with capacity to deliver cost-efficient and science-based solutions, able to transform and handle existing data from (often) data silos across various public agencies servers. Such a system developed under the umbrella of actionable intelligence methodology will leverage the true potential of the chosen approach.

Theoretically, this approach can be deployed in any public or business sector and integrate data of many related sectors into new data intelligence entity in a fully automated and straightforward manner. It is perceived as an analytic tool for spotting trends rather than performing analytics on the fly, with stronger emphasis on periodical loads of existing or even empirical data.

It can be used, by the same token, for effective provisioning of public assistance to certain targeted groups (immigrants, people from isolated communities, etc.) to alleviate the effects of social disadvantages.

However, it could also serve as a base for creation of enriched predictive models if used in middle-term strategies.

In a data-driven world, actionable intelligence proves to be a prospective methodology with expected impact on social perception, delivering benefits both practical and financial to the public sector and to the community in general.

# References

Adriana Lleras-Muney, The Relationship Between Education and Adult Mortality in the United States, 2004

Agathe Merceron, Educational Data Mining - a Case Study, 2006

Antonio Cordella et al., Information systems in the public sector - The e-Government enactment Framework, 2010

Barry Lawrence et al., Best Practices for Boosting Profitability, 2011

Barry Lawrence et al., Customer Stratification - Understanding Customer Profitability, 2011

Bhavani Thuraisingham, Data Mining for Counter Terrorism, 2004

Bo Sundgren, - What is a Public IS, 2007

Brijesh Kumar Baradwaj, Mining Educational Data to Analyze Students Performance, 2011

Bryson et al., Design and Implementation of Cross-Sector Collaborations, 2006

Bryson et al., Leadership and the Creation of Cross-Sector Regimes ,2005

Budhraja, Trends and Innovation with Service Reuse, Cloud and Big Data, 2012

Chris Livesey, Social Distribution of Crime, 2010

Culhane, Actionable Intelligance For Social Policy, 2015

Culhane, Connecting the Dots, 2010

Culhane, The promise of integrated data systems for social policy reform, 2016

Daniela Grudinschi, Management Challenges in Cross-Sector Collaboration - Elderly Care Case Study, 2013

Duncan Cleary, Predictive Analytics in the Public Sector, 2012

DataSparx, Actionable Intelligence Paper, 2016

Fantuzzo, Actionable Intelligence, 2015

Fantuzzo, Kids Integrated Data System, 2009

Frank Hopfgartner, Smart Information Systems Computational Intelligence for RealLife Applications, 2015

Groves et al., the pathways could account for, 2013

Heiner Stuckenschmidt et al., A Study In User-Centric Data Integration, 2012

John S. Hollywood et al., PredictivePolicing, 2011

Jordan & King, Building Effective Systems for Young Children and Families, 2015

Kirk Emerson et al., Integrative Framework for Collaborative Government, 2011

Lichun Chiang et al., Information Integration to Create an Infrastructure, 2007

Lu Zhikui, Understanding Cross-Sector Collaboration In E-Goverment, 2010

Maria Cristina Limlingan, Integrated Data Systems, 2015

Michael Barber et al., Global trends affecting public sector, 2007

Michael E. Milakovich, Anticipatory Government, 2013

Michael Lock, BI In The Public Sector, 2012

Missy Cochenour, Early Childhood Integrated Data System, 2015

Mohammed T. Al-Sudairy et al., Semantic Data Integration, 2011

Ojo-Janowski-Estevez , Whole-of-government approach to IT, 2011

Osmar R. Zaïane, Principles of Knowledge Discovery in Databases, 1999

Pablo Fajnzylber, Inequality and Violent Crime, 2002

Peter Dröll, Public Sector Innovation - Towards New Architecture, 2013

Petter Gottschalk, e-Government Interoperability, 2009

Prashant Kumar, An Overview of Architectures and Techniques for Integrated Data Systems (IDS) Implementation , 2015

Raul Abril ,The European Interoperability Reference Architecture, 2015

Roberta Nannucci et al., Converting Online Public Legal Information Into Knowledge, 2006

Shawn P. McCarthy, Data integration and homeland security, 2004

Thomson & Perry, Collaboration Processes - Inside the Black Box, 2006.pdf

United Nations E-Government Survey 2012, Chapter 3, Taking a whole of government approach, 2012

United Nations E-Government Survey 2014, 2014

Velamala Ranga Rao, A Framework for e-Government Data Mining Applications, 2014

Walter Bailey, Integrated State Data Systems, 2015

Zhang Nan, Brysons Framework Impact, 2013

Zhang Nan, Cross-Sector Collaboration and Information, 2012