1 DataSTAT Hub: a tool for the automatic collection of administrative data and metadata to produce official statistics

Alessandro Capezzuoli, Emanuela Recchini ([1])

1.1. Introduction

The need for relevant, reliable and even more timely statistical data to support decision making process and scientific research has contributed to a growing demand for new statistical information to best analyze and rule, at various levels, the deep social, economic and environmental changes occurred at regional and global scale. Official statistics, characterized by the highest quality possible inasmuch as they are produced in compliance with the United Nations Fundamental Principles of Official Statistics and the European Statistics Code of Practice, are best suited to meet this need.

There is worldwide recognition of the increasing role played by administrative data and metadata in the production of prompt and more disaggregated statistics at higher frequencies than traditional survey data. They offer further information on a wide range of issues, including some which cannot be answered cost-effectively from survey data.

The efficient use of all available information to produce timely, accurate and high quality statistics is a challenge for National Statistical Offices (NSOs), which are even more committed to developing methods and suitable tools for the production, collection, standardization and integration of different types of statistical data. The bringing together of information from different sources makes it possible to fill information gaps or provide insights which cannot be gleaned from unlinked data and to improve the knowledge and understanding about specific phenomena.

In this chapter, a short review on the collection of administrative data and metadata for statistical purposes, with particular reference to Istat experience, is given. Then, some critical issues and possible technological solutions related to data collection processes are considered. including what is offered by the Web and is at the basis of DataSTAT Hub, a tool for collection, and release of data and metadata.In this respect, an overview of technology, model and architecture used to create DataSTAT Hub permits to describe its main features. A specific focus is put on an example of application of this hub to classifications of official statistics.

[1] Italian National Institute of Statistics (Istat).
The views expressed in this paper are solely those of the authors and do not involve the responsibility of Istat. Addresses for correspondence: alessandro.capezzuoli@istat.it and emanuela.recchini@istat.it.

1.2. Data collection and official statistics

1.2.1. Data collection process

Accurate data collection is essential in all fields of study, from social and physical sciences to economy, politics, environment, etc.: it permits to answer research questions, validate hypotheses and evaluate outcomes.

The production of statistics based on administrative data and metadata from different sources is closely related to the methods and techniques of collection, integration and management of archives.

Problems related to automatic data collection are numerous as they involve the production and standardization of outputs, in order to: make them usable by web applications; be stored in a database; be connected (record linkage); be processed by statistical software and/or visualized within *ad hoc* created web platforms.

In this respect, it is worth keeping in mind that in a data collection process both the data collector and data supplier have specific tasks; the main ones are summarized in the box below.

Main tasks of data collector and data supplier

DATA COLLECTOR

- manages data requests
- defines methods and standards
- monitors and handles reminders
- stores data and metadata
- standardizes and disseminates

DATA SUPPLIER

- receives data requests
- elaborates data requests
- prepares data to be sent
- sends data to data collector

In order to ensure compliance with the United Nations Fundamental Principles of Official Statistics and the European Statistics Code of Practice, particularly as far as the quality and reliability of the data collected by NSOs for statistical purposes is concerned, some aspects related to the different sources of data have to be considered. Data suppliers may provide:

- administrative registers
- synthetic indicators
- metadata
- Big Data

Administrative registers contain data collected by data suppliers with different criteria that often data collectors do not know.

In order to meet quality standards for the data, NSOs change an administrative register into a statistical register, through check, processing, standardization and normalization algorithms. This process ensures the usability of data for statistical purposes and the production of synthetic indicators also by linking records from different sources with each other.

Among the most recent activities carried out by NSOs, Big Data collection plays a key role in producing statistics and monitoring numerous phenomena. Big Data can be collected through two main channels: data supply from institutional and non-institutional agencies (Agency of revenue, telephone companies, Police Headquarters, etc.) and data scraping, namely a technique to extract data published on websites (e.g. online booking sites, advertising sites, etc.) through data capture and storage procedures.

Each data collection process presents critical issues like handling data requests and reminders, complex IT infrastructure, a burden for data supplier, human resources for transactions

management. Among the different solutions, there is the possibility to collect data through File Transfer Protocol or upload data through an *ad hoc* website to handle reminders and data supply requests. However, these solutions do not permit to automate the process.

Notably in the case of Big Data, IT architectures used for data collection just mentioned above are not very suitable to store large volumes of data and interface with applications for data mining and/or data analysis. Solutions requiring data standardization (e.g. SDMX, JSON-Stat, etc.) can be very onerous for data suppliers, since they use structured databases based in most cases on *ad hoc* schemes.

The World Wide Web offers a possible solution. In fact, HTTP (Hypertext Transfer Protocol), i.e. the set of rules for transferring files on the Web, can be conveniently used for data collection and data exchange. This will be more fully explained in paragraph 1.3.2.

1.2.2. Administrative data collected by Istat

At present, the exploitation of administrative data and metadata for statistical purposes is a normal practice for a large number of NSOs. This improves the quality of statistical outputs, eliminates process redundancies, reduces the statistical burden on respondents and minimizes costs.

According to the provisions of the Italian Digital Administration Code[2], before proceeding to the collection of new data, public administrations are required to verify whether the information they need can be acquired through access to information already in the possession of other public authorities or public bodies.

Technical options for data usability

- web access through the website of the supplier institution or an *ad hoc* thematic website
- **interoperability among public administrations for data collection and data integration**
- the user can process data collected exclusively for the pursuit of its institutional goals; data transfer from one information system to another does not change data ownership
- the transfer of a data from an information system to another does not change the ownership of the given

(Guidelines for the drafting of conventions on the usability Public Administrations data; Legislative Decree n. 82/2005, commonly referred to as the “Digital Administration Code”, modified by the Legislative Decree n. 235/2010)

The Italian National Institute of Statistics (Istat) collects and manages a large volume of administrative data and metadata from different sources, among which: Italian Agency of Revenue; Bank of Italy; Ministries; Social Security and government Institutions. From 2009 to 2016, administrative data sets supplied to Istat have trebled. In this respect, the need to engineer processes in order to automate and manage collection and release of administrative data and metadata is increasingly urgent. Data collected by Istat are very different from each other in type, content and structure. Administrative registers may include statistical data, micro data, geographic data, synthetic indicators and many other types of data and metadata.

The main channel used by Istat for the collection of administrative registers from public and private institutions is the ARCAM website[3]. Through this portal it is possible to manage data requests, remainders and upload files containing administrative data and metadata. ARCAM meets the need of managing centralized repositories for data collection in compliance with legislation on the confidentiality of the data.

Another organizational and technological environment to share, integrate and disseminate data and

[2] For further details, see the Italian Legislative Decree no. 82/2005, subsequently integrated and amended by Supplementary Provisions and the corrective Legislative Decree no. 159/2006.
[3] For further details on ARCAM, see: https://arcam.istat.it/arcam/

metadata within the Italian National Statistical System (Sistan) is Sistan Hub[4]. This informative system, in line with the objectives of the Digital Agenda, supports the semantic interoperability among different Institutions, facilitates data searching and uses SDMX international standard as main requirement for their representation. Although it adapts to the different needs of NSOs, Sistan Hub requires an investment in human resources for data mapping and creation of the single exit point and this solution is not very suitable for Big Data management and for data suppliers who not always have a thorough understanding of SDMX standard.

The imposition of standards and the creation of IT infrastructures for data-exchange would be very onerous in terms of costs and time of execution for data suppliers.

1.3. DataSTAT Hub for automatic collection and release of data and metadata

1.3.1. What is DataSTAT Hub?

DataSTAT Hub is a suitable and easy tool for collection, standardization, integration and release of data and metadata. It automates these processes through HTTP and a model that simplifies the structure of data.

By taking advantage of the potential offered by HTTP, DataSTAT Hub can be used through two different architectures: star or centralized. The former implies that each microservice (hub node) is automatically populated by data supplier through a set of query strings and can be accessed in reading by the central institution that performs data collection. The latter architecture implies the automatic population of the central hub that interfaces with the various institutions through the just mentioned query strings and allows the data collector to store data and metadata for example in a NoSQL database (Cassandra) using the key-value data model for their representation.

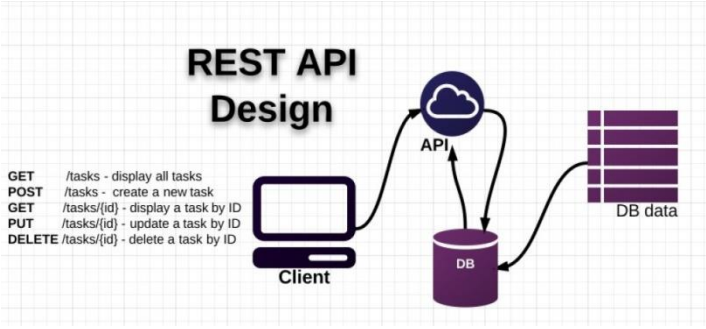
DataSTAT Hub permits to standardize outputs in various formats (XML, JSON, CSV) and models (JSON-stat, SDMX, DDI).

1.3.2. DataSTAT Hub technology

As mentioned above, HTTP can be conveniently used for data collection and exchange. It is a request/response protocol based on the client-server architecture.

A set of guidelines to use HTTP is defined by REST (Representational State Transfer), an architecture style for designing networked applications. REST architecture permits to separate relational databases from the client through an API (Application Programming Interface), which exploits HTTP to transmit data and exchange information (Figure 1).

Figure 1: REST API architecture



[4] For further details on Sistan Hub, see: http://sistanhub.istat.it/hub

REST allows data supplier to perform CRUD operations(Create, Read, Update and Delete) with a logic similar to that used on any SQL database (Figure 2).

Figure 2: HTTP, CRUD and SQL operations

HTTP	CRUD	SQL	DESCRIPTION
POST	CREATE	INSERT	Create or add a new resource
GET	READ	SELECT	Read, retrieve, search, or view existing resources
PUT	UPDATE	UPDATE	Update or edit existing resources
DELETE	DELETE	DELETE	Delete/deactivate/remove existing resources

1.3.3. DataSTAT Hub model

The creation of a model to collect different types of data and metadata requires important features to be taken into consideration, in order to simplify, facilitate and generalize the process, among which:

- *unstructured data* – a model collecting data in their essence (key-value) is more convenient and immediate than defining multiple standards for data representation;
- *scalability* – a highly extensible architecture is needed, in case of possible conceptual/architectural future upgrade;
- *intuitive schema* – the model should be easily applied by data suppliers, without resorting to complex studies of any imposed standard;
- *big-data-oriented architecture* – storage is closely linked to the tools used for semantic search, data analysis and data visualization. Elasticsearch, Hadhoop, Solr, Cassandra provide a complete integrated environment for managing them.

When dealing with highly heterogeneous data, it is recommended to use a model to represent them in their simplest form: a key-value pair. The format that is better suited to use HTTP and transmit data objects consisting of key-value pairs is JSON (JavaScript Object Notation) to which different models for data representation can be associated. This storage model, at the basis of DataSTAT Hub, is the same used by some NoSQL databases (for example, Cassandra), well suited for a big data processing architecture.

DataSTAT Hub model is characterized by a five-level recursive structure based on key-value logic.

SKEY Statistical Key Value data model

```
{
"keyspace" :
   {
    "columnfamily" :
     {
       "rowkey" :
         {
         "supercolumn" :  { "column name" : "column value"}
         }
     }
   }
}
```

Each data supplier has and autonomously manages one specific keyspace.

There are two column family types in this model: data and metadata.

The number of super columns, namely aggregates of columns, is defined on the basis of data supplier's needs.

Columns contain key-value pairs.

1.3.3. DataSTAT Hub architecture

DataSTAT Hub uses Elasticsearch, an open source search engine that can be conveniently used for collection and release of data[5]. Through Elasticsearch it is possible to index and map documents/data through query strings to be sent via HTTP in JSON format.

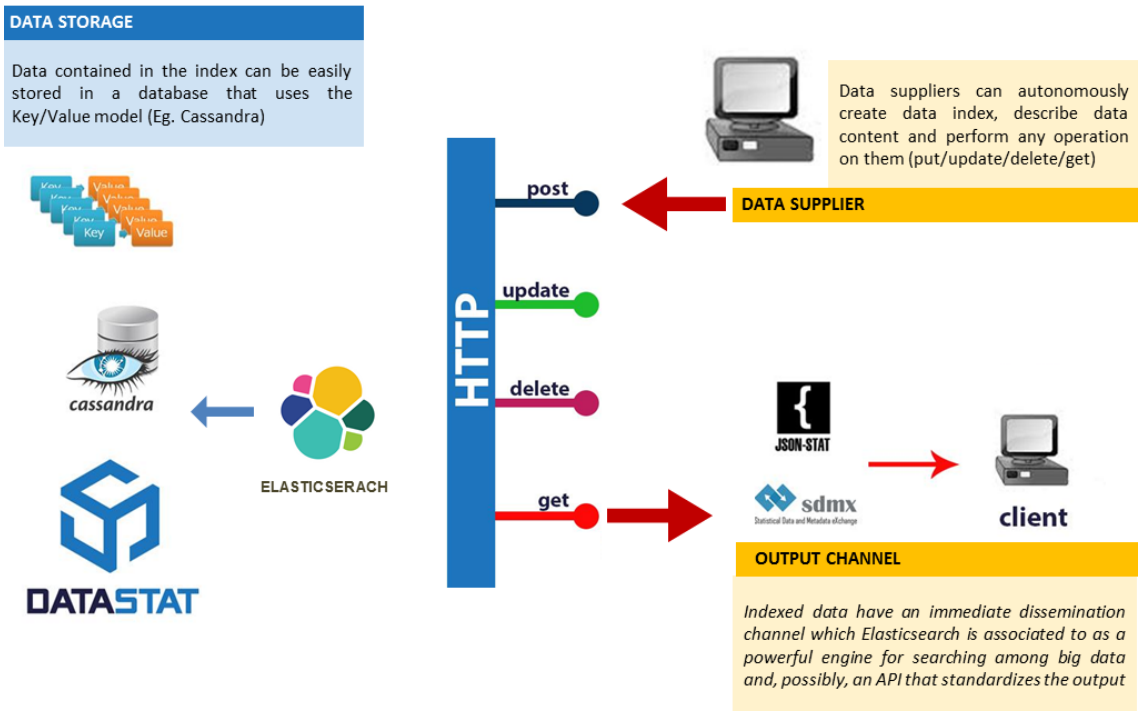
ELASTICSEARCH MAIN FEATURES*

- DOCUMENT: most entities or objects in most applications can be serialized into a JSON object, with keys and values. A key is the name of a field or property, and a value can be a string, a number, a Boolean, another object, an array of values, or some other specialized type such as a string representing a date or an object representing a geolocation.
- INDEX/TYPE: documents are indexed – stored and made searchable – by using the index API, which uniquely identifies the document.
- MAPPING: is the process of defining how a document and the fields it contains are stored and indexed.

* For further details, see: https://www.elastic.co

The phases of data collection and release process through DataSTAT Hub are illustrated in Figure 3.

Figure 3: DataSTAT Hub architecture and data collection and release process



[5]For further details on Elasticsearch, see:https://www.elastic.co/

1.4. Statisticlass: DataSTAT Hub applied to statistical classifications

1.4.1. Acquisition and management of statistical classifications

DataSTAT Hub is well suited for the solution of some critical issues related to the use of statistical classifications in different fields (surveys, administrative registers, information systems, etc.), such as:

- acquisition, storage, management and updates of classifications;
- multilingual semantic search for coding;
- sharing and dissemination of coding tools.

Each of these issues has been effectively resolved within the project Statisticlass (http://www.statisticlass.eu) thanks to the exploitation of DataSTAT Hub.

Statistical classifications have a generalizable logical structure, described within the Generic Statistical Information Model (GSIM[6]), which provides the organization of content in a complex and structured architecture (see box below on GSIM). Acquisition, semantic search and dissemination of classification items can be managed effectively through DataSTAT Hub.

To this end, it is possible to exploit a very simple JSON object, to which then associate the metadata related to the classification (family, serie, level, etc.).

JSON OBJECT

```
{
  "code":"x.x.x",
"name":"NAME",
  "description":"DESCRIPTION"
}
```

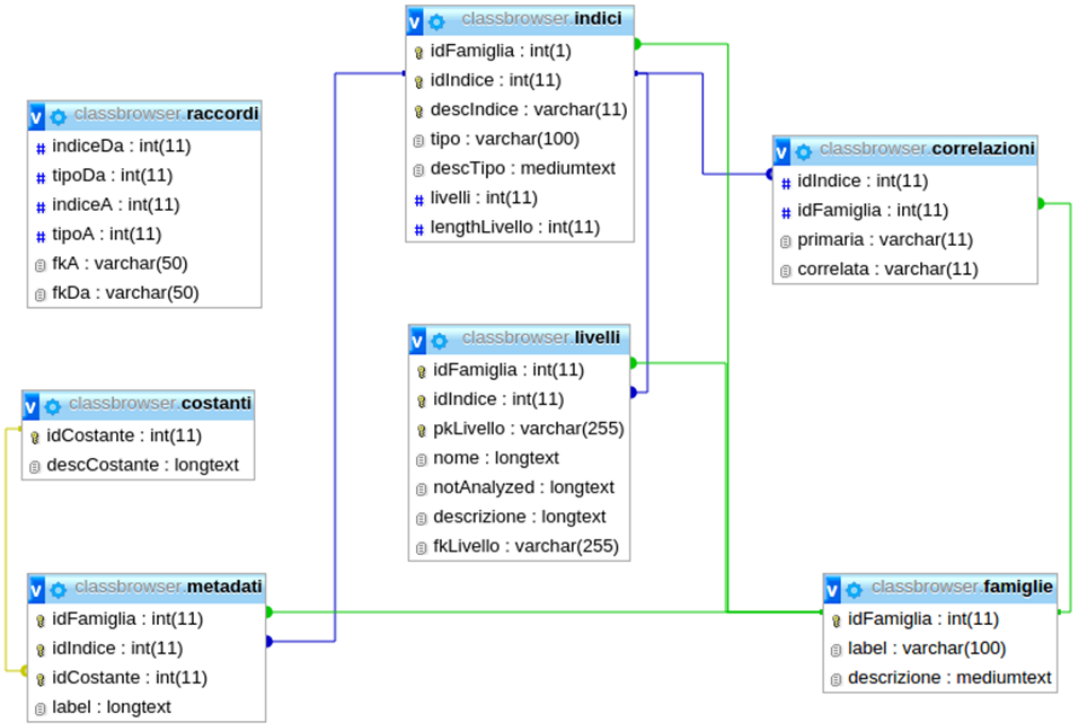
GSIM

- Classification Family
- Classification Series
- Level
- Correspondence Table
- Classification Index
- Classification Item
- Classification index entry

PUT and GET methods (mentioned in Figure 2) of DataSTAT Hub permit an easy acquisition of classification items which can then be organized through *ad hoc* procedures, on the basis of GSIM model, and stored into a relational database (Figure 4).

[6]For further details on GSIM, see: http://www1.unece.org

Figure 4: Statisticlass entity–relationship model



1.4.2. Semantic search and coding

Textual search is a very popular technique for users who seek information on the web. It does not require any special skill users have already acquired through surfing the web and it is also suitable to search within statistical classifications and facilitate coding. The most common problem related to semantic searches within taxonomies concerns false-positive and false negative results. The search is usually done through SQL queries allowing users to perform two types of operations: "exact match" and "full text". String parsing algorithms can be associated to the SQL queries.

The main critical issues related to textual searches imply the management of:

- singulars and plurals
- stopwords, namely words irrelevant for research that depend on the dictionary and language
- male and female terms
- differences between a one-word search or a sentence search
- case sensitive
- misspellings

As well as enabling data and metadata collection through HTTP methods, DataSTAT Hub uses Elasticsearch, which is a powerful search engine based on Lucene library and providing numerous RESTful services useful to perform complex semantic searches through JSON query.

A statistical classification can be indexed within Elasticsearch to perform complex and differentiated textual searches through DSL (Domain Specific Language) in JSON format. This solution permits to simplify the formulation of complicated SQL queries and makes the search system from any programming language usable.

Elasticsearch allows users to manipulate large volumes of data thanks to an internal document management, completely independent from relational databases, and the opportunity to create distributed cluster. The creation of index and mapping related to the "occupation" family and the

"name" field of the International Standard Classification of Occupations (ISCO-08) are synthesized in the following boxes.

INDEX

Creation of the index related to the "occupation" classification family

```
PUT occupation
{
   "settings" : {
      "number_of_shards" : 3,
      "number_of_replicas" : 2
   }
}'
```

MAPPING

```
PUT occupation
{
  ""mappings": {
   "ISCO08": {
    "properties": {
     "name": {   "type": "string",
     "analyzer": "standard"
     }
    }
   }
```

The analyzer is the tool to build queries through a JSON array. The DSL language is made up of: the tokenizer, namely a system dividing text into individual tokens; a set of filters through which users can fix the criteria for making the search string parsing. The result of a search is an output in JSON format. The items are listed on the basis of a score assigned in function of the discriminating power of the searched string.

The search engine of DataSTAT Hub applied to Statisticlass implies different steps for different levels of precision. The first operation performed by the search system consists of the identification of the exact matching of the string. The results obtained are enriched by a more detailed analyzer, which provides a sophisticated stemming algorithm. This algorithm is combined with a dictionary for stopwords management. Another precision level includes the correction of the syntax of entered strings. In this way, a very precise result is achieved, false positive and negative results are reduced and users can make use of different channels for dissemination and sharing of classifications.

It is worth pointing out that such search systems can be used with different levels of precision to any statistical classification.

1.4.3. Sharing and dissemination of classifications and search engines

DataSTAT Hub provides a large set of tools for sharing and dissemination of statistical classifications and search engines. In this respect, an immediate channel is the REST web service provided by Elasticsearch. The web service output, in JSON format, is obtained through a query string (see box below). This solution is well suited to the development of web applications that need an efficient, uniform and high level of customization coding system.

JSON OUTPUT

QUERY STRING


```
{  "took":18,
   "timed_out":false,
   "_shards":{  "total":5, "successful":5,"failed":0 },
   "hits":{ "total":1,  "max_score":3.1464005,
      "hits":[
         {   "_index":"isco08",
            "_type":"isco08",
            "_id":"AVe31A1PIZ7xXoOaTaQB",
            "_score":3.1464005,
            "_source":{
               "pkLivello":"2111",
               "nome":"Physicists and astronomers",
               "fkLivello":"211"
            },
            "highlight":{
               "nome":[ "<em>Physicists<\/em> and
astronomers"  ]
            }
         }
      ]
   }
}
```


Figure 5 summarizes the results of the search within Statisticlass website (http://www.statisticlass.eu).

Figure 5: Statisticlass web interface

The use of web services requires high-level IT skills, therefore it is reserved for a selected group of users (mainly programmers or web developer).

Easy to use widgets have been developed to include coding systems in Statisticlass within web questionnaires or web applications.

A widget is a script that can be included within a web application through a simple "cut and paste" (almost as it happens for multimedia content included in social networks).

Istat experience in using this methodology has been very satisfactory. The coding systems related to the main statistical classifications (ISCO, NACE, ISCED, COFOG, COICOP) were included in several Istat surveys ("Labour Force Survey", multi-purpose survey "Aspects of daily life", "Consumer prices", etc.) and Information system on occupation[7]

1.5. Concluding remarks

The importance to complement existing data derived from traditional surveys with those from administrative sources is worldwide recognized. They offer further information on a wide range of issues, including some which cannot be answered cost-effectively from survey data, and make it possible to improve the knowledge and understanding of specific phenomena.

What do problems related to data collection involve? Summarizing: the production and standardization of outputs to be stored in a database, connected, processed by statistical software and/or visualized within *ad hoc* created web platforms.

[7]For further details on the Information system on occupation, see: http://professionioccupazione.isfol.it

DataSTAT Hub is a suitable and easy tool for automated collection, standardization, integration and release of administrative data and metadata. It permits to reduce the burden on users, because it does not require the knowledge of the internal data base since the updating is performed through the HTTP query strings and can be used with any programming language; once created, the procedure will be used for each next data supply.

Differently from previously mentioned data collection systems such as File Transfer Protocol and ARCAM, DataSTAT Hub can support all phases of statistical process. In addition to data and metadata collection, this informative system permits to connect to data mining and data analysis applications (e.g. Hadoop) and visualize and disseminate statistical data (Kibana, Spago BI, Micro Strategy, etc.).

Instead, for the purpose of data treatment and validation, the most common techniques, implying the population of traditional data bases linked to the most usual statistical software (SAS, R, SPSS, etc.), can be used.

Not only do users have a reduction of burden but also a reduction of costs in terms of employment of human resources – for organizational, bureaucratic and IT management – and time-saving, by permitting to overcome some critical issues related to the use of administrative data, including those connected with confidentiality of the data. Moreover, DataSTAT Hub can be effectively applied to statistical classifications in order to facilitate semantic search, dissemination and sharing.

It is a user-friendly tool developed by making use of open source technologies (PHP, MySQL, Cassandra) and can be conveniently shared among NSOs, while it is extensible to any other institution interested in the automatic collection and integration of administrative data and metadata.

References

Gormley C., Zachary T. (2015), *Elasticsearch: The Definitive Guide. A Distributed Real-Time Search and Analytics Engine*, O'Reilly Media.

Masse M. (2011), *REST API Design Rulebook*, O'Reilly Media.

Olsen W. (2012), *Data Collection. Key Debates and Methods in Social Research*, SAGE Publishing.