Mediating Open Data Consumption – Identifying Story Patterns for Linked Open Statistical Data

Maciej Janowski
Insight Centre for Data Analytics, National University of Ireland, Galway
maciej.janowski@insight-centre.org

Adegboyega Ojo
Insight Centre for Data Analytics, National University of Ireland, Galway
adegboyega.ojo@deri.org

Edward Curry
Insight Centre for Data Analytics, National University of Ireland, Galway
ed.curry@insight-centre.org

Lukasz Porwol
Insight Centre for Data Analytics, National University of Ireland, Galway
lukasz.porwol@insight-centre.org

ABSTRACT

Statistical data account for a very large proportion of data published on open data platforms. This category of data are which are often of high quality, value and public interest; are gradually being published as 5-star linked open statistical data or data cubes (LOSD) for easy integration and cross-border comparability. However, publishing open data as linked data (i.e. graph oriented) significantly increases the technical skill requirements for end-user consumption. We address this problem by mediating the exploration and analysis of LOSD published on open data platforms through the use of data stories. After providing the requisite background information on LOSD, we identified data story patterns from extant literature and show how these patterns can be employed in analysing LOSD. Subsequently, we provide a case study to illustrate the use of these data story patterns as an end-user domain-specific language to explore and analyse LOSD. We argue that using data stories for exploring and analysing on open data platforms has the potential to significantly increase the adoption and use of (linked) open data.

CCS CONCEPTS

• Applied computing → Computers in other domains → Computing in government → E-government

KEYWORDS

Data storytelling patterns, Linked Open Statistical Data, Open Data Platforms, Data Cube Vocabulary

ACM Reference format:

M. Janowski, A. Ojo, E. Curry, L. Porwol. 2019. Mediating Open Data Consumption – Identifying Story Patterns for Linked Open Statistical Data. In *Proceedings of the 12th International Conference on Theory and Practice of Electronic Governance (ICEGOV2019), Melbourne, VIC, Australia, April 3-5, 2019*, 8 pages. https://doi.org/10.1145/3326365.3326386

1. INTRODUCTION

There is wide acknowledgement of the wide availability of open data on open government data portals and portals maintained by international organizations such as the World Bank. However, there is equally growing consensus that the use of these datasets are still relatively limited [24]. Some of the identified barriers include low data literacy, poor data and metadata quality non user-friendliness of portals. In addition, current generation of open data platforms generally provide limited support for data analysis and sense-making to ordinary or non-technical end-users [25].

Lately, ideas for extending open data platform with data story features have been considered. According to [4], open data platforms with support for data storytelling are required to among other things support the discovery of relevant data, provide assistance in wrangling data and support the generation of insights from stories. To enable the generation of insights from data published on open data platforms, tools for data analysis and visualisations are required.

Recent studies such as [26] also show that both small and large newsrooms are employing data story telling using publicly available data including open data in their data journalism practices. However, apart from [4], we are unaware of investigations linking data stories with data published on open data platform. We address this gap in this paper by examining how data stories can be used to mediate the use of linked open statistical data available on open data platforms.

Statistical data are often high quality data due to the relatively mature data practices in national or regional statistical institutions. In addition, this category of data are of high public

ICEGOV2019, April 3–5, 2019, Melbourne, VIC, Australia
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6644-1/19/04…$15.00
https://doi.org/10.1145/3326365.3326386

interest. For instance in Ireland, statistical data comprise 49% (4358 datasets) out of the currently published 8810 datasets on data.gov.ie. These datasets have also been viewed over 70,912 times. Increasingly, statistical data are published in linked data (LOSD) formats with specific shared vocabularies to facilitate the integration with other datasets and also cross-border comparability (e.g. with the European Union). However, while linked data formats are considered the holy-grail (so-called 5-star data), they significantly raise the barrier to access and understandability for both ordinary users and in fact traditional web developers [32]. Thus, we consider mediating the consumption and exploration of LOSD through mechanisms including data storytelling compelling and necessary for widespread use of this category of high-quality dataset.

The rest of this paper provides some background to LOSD and data storytelling before presenting a framework which aligns data story patterns to LOSD (or RDF Cube) operations.

2. LINKED OPEN STATISTICAL DATA

2.1. Linked Data

Linked Data in general is a way of publishing structured data across the Web, which allows interlinking resources between each other. Such approach allow integrating data from different sources into one big network of related records. When a user looks for data from a specific field he/she has to browse pages one after another trying to explore all data possible [13]. Even if there is automated process of data search it relies on APIs provided by each data center, what conducts slicing the Web into separate data silos[3]. This problem is addressed with Tim Berners-Lee's concept of standardizing Web technologies to allow generating links between resources on level of data. Such connection between data sources allow automatic crawling and retrieving all of the possible data interesting for user without a need to browse by himself/herself. The guideline for such practices is called Linked Data Principles[19]: 1) use URIs(2) as names for things, 2) use HTTP URIs so that people can look up those names; 3) when someone looks up a URI, provide useful information, using standards (RDF, SPARQL) and 4) include links to other URIs, so that users can discover more things.

The whole idea of Linked Data is based on RDF (Resource Description Framework). Based on [30] design goals were as it states: having a simple data model; having formal semantics and provable inference; using an extensible URI-based vocabulary; using an XML-based syntax; supporting use of XML schema datatypes; allowing anyone to make statements about any resource.

Any RDF expression is expressed as a collection of triples :subject, predicate and object. A group of such triples is called a RDF Graph – each subject point out to object based on a given predicate (property of relationship). For example RDF triples can represent two families (subject and object) which are neighbours living on the same street (predicate – property of their relationship).

2.2. Linked Statistical Data

Linked Statistical Data can be defined as statistical data published on the Web following Linked Data principles. To publish reliable linked statistical data, the process includes:

- Refining Data – checking for misspelling, empty fields, redundant whitespaces
- Data Transformation – from structural files into RDF
- Defining the RDF Structure – developing structural data models based on ontologies and vocabularies
- Designing URI – preparation of URI scheme
- Creation of metadata – background information about the data.

Publishing government agency data following Linked Data principles enables cross-agency data integration based on agreed semantic standards and assets. Such interlinked volume of open government data provide better support for better decision/policy making, better service delivery, joined-up thinking across government.

The Data Cube Vocabulary presented in [33] provides the concepts for describing and representing Linked Open Statistical Data as multi-dimensional data model. The vocabulary is based on SDMX ISO standard describing rules of statistical data exchange. The Data Cube Vocabulary is based on the following RDF vocabularies:

- SKOS (Simple Knowledge Organization System) for concept schemes
- SCOVO (Statistical Core Vocabulary) for core statistical structures
- Dublin Core Terms for metadata
- VoiD (Vocabulary of Interlinked Datasets) for data access
- FOAF for agents
- ORG for organizations

However, while standard Web technologies such as HTTP, RDF and URIs and standard vocabularies such as RDF data cube, SKOS and XKOS are used for LOSD publishing, different portals often adopt different practices for using these standards [32]. According to the same source, different practices are often adopted for the definition of popular measures and popular dimension (i.e. time, geography) along with their code lists. This implies the needs for government-wide standardisation and agreement on these semantic assets and related practices.

2.3. Application of Linked Open Statistical Data

Linked Open Statistical Data allow exploration across interlinked data instead of getting data by browsing each data source manually. The rest of the section describes recent research about LOSD analysis and visualizations.

Authors of [22] examine usage of LOSD for urban-policy making. Officials, who are responsible for their specific sectors,

2 Unified Resource Identifier – identifier of the resource compatible with given scheme

while creating policies they might not fully understand problems and issues of other sectors. Such misunderstanding when the policy has been applied can exacerbate the problem rather than solving it. LOSD is used to create shared knowledge and understanding. What is more, analysis and visualisations made with Linked Data help policy-making group to develop shared view of the problem.

Given that LOSD may too complex for regular open data platform users, researchers are motivated to create easy-to-use tools for visualizations of Linked Data. Authors of [23] created a number of tools including a prototype visualizer called "LODViewer", which allows to illustrate how different RDF sources including SPARQL endpoints. Implementation was based on LDVM (Linked Data Visualization Model) proposed in paper [5], which supports visualisations, analytical operation on Linked Data. By implementing the tool in JavaScript LODViewer. The complexity of using the tools is reduced it is easier ease-of-use pilot concept to encourage normal people to gain more interest in Semantic Web and the idea of Linked Data

RDF graph generated as an output of a given query is illegible due to the amount of connections and lack of organization. Authors of [16] propose an approach to make the graph more reader-friendly by three steps:

- Graph simplification (reduce redundancy)
- Triple ranking (triples sorting sourced from discrete levels of knowledge)
- Property selection (customize graph making it compatible with your interests)

Those operations assumes that RDF data can be described as knowledge units, on which statistical calculations can be held to deliver easy-to-understand graph to users. Graph representation of data might be difficult for end-users to retrieve knowledge from. it – that is why simplicity of those graphs is required to make understanding as user-friendly as possible.

Complexity of Linked Data processing is the biggest issue with end users. Authors of [35] propose decomposition of complex Linked Data technology into small widgets, which might be connected with each other, doing simple tasks. Such environment will eliminate necessity of technical expertise in Linked Data – simple operations will allow end-users to create applications satisfying their needs. Whole research is based on Linked Widgets – single purpose program that operates on Linked Data. They define three categories of such widgets:

- Data widgets – data providers for other widget types (processes data to required format)
- View widgets – produces visualizations as an output based on given parameters
- Process widgets – provides required processing of the data (filtering, merging, etc.)

Such platform conducts user-friendly environment for exploring Open Government Data without need of technical expertise in data processing.

Linked Open Statistical Data might also be used as a support in public administration [15]. Examples describing the utility of LOSD in government data:

- The Flemish Government - Comparison of pollution levels between regions and companies will give citizens information what are emission of pollutants, which may conduct creating better policies solving problem with pollution
- The Estonian - Ministry of Economic Affairs - Portal with detailed information about nearby area will ease search for such knowledge through many websites. Citizens interested in specific information about area, in which they live/to which they move in, will be found on one portal.
- The Greek Ministry of Interior
  Usage of LOSD allows simplification of management process of vehicles monitored by mentioned ministry. Datasets describing status and needs comes from different sources and have to streamlined to make it easier for logistics staff to manage their fleet in more efficient way.
- The Irish Marine Institute
  In Ireland due to diverse coastline there many search and rescue operations held every year. Historical data about such situations may allow search and rescue staff to detect key areas, in which missing people might be found, on specific section of coastline. The system will visualize historical data on a map to reduce operational time of operations – search and rescue staff know where to look first.
- The Lithuanian Ministry of Economy
  Entrepreneurs from Lithuania would need to know about investment opportunities across country. They need to know about if they interesting location has any potential. Getting such information is time- and resource-consuming process. Creating a portal with all information entrepreneurs need will simplify business expansion process.
- UK Trafford Council
  Using LOSD in Job Centres management process will allow easier handling of people logistics (allow people from different Job Centres find job easier) – showing where are greater needs for employees. That can be also held without LOSD but it will consume a lot of time resources of different Job Centres. Applying LOSD in that process reduces cost and increase performance to make citizen get served faster.

With increasing amount of information being processed, data mining becomes indispensable. Research paper [28] propose a framework to implement data mining into LOSD. They developed RapidMiner extension compatible with Linked Data. Such platform can be used form:

- Text Classification – extraction of events from Wikipedia, detecting emergency situation based on Tweets.
- Explaining Statistics – providing background knowledge for statistical operations.

Linked data mining deliver many opportunities but also many challenges to deal with (for example: variety of data representations). LOSD in data mining might be described as background knowledge provider to enrich experience and understanding of given situation.

3. ANALYSING LOSD

3.1. OLAP Data Model

Every company produces uncountable amount of data about their actions (transactions, sales, marketing, etc.). That data stored in databases is analysed to detect patterns and errors in past operations in order to prevent repeating mistakes and support decisions for future actions. Data warehousing and On-line analytical processing are the most crucial elements in system backing up decision making. Data warehouses store past operational data, which allow to detect patterns and trends (for example: in sales), and OLAP technologies provide tools for analysing large quantities of information. [20].

Formulated by [9] OLAP analysis are applied to business analysis to detect anomalies in multidimensional dataset storing all the information about business operations. Data is organised as a set of dimensions (as column names in database table), facts and measures (numerical values), with which set of dimensions is related creating a meaning to exposed number. Such data organisation method might be too complicated to analyse as a whole. That is why, OLAP multidimensional model allow several operations simplifying data exploration and reducing of its' complexity[7]:

- Roll-up – reviewing data on different levels of granularity
- Drill-down – dimensions inspected at lower hierarchy levels
- Slice – extracting data with one limited dimension.
- Dice – limiting multiple dimensions and extracting "subcube".
- Pivot – rearranging data based multi-dimensional aggregation

On-line Analytical Processing (OLAP) is about exploration of multidimensional data cube and discovering interesting data patterns and anomalies. Such exploration is handled by filtering one or more dimensions to create different presentations of data. However, authors of [18] claim that traditional OLAP operations do not allow to discover all important patterns possible. They propose possibility of new dimension aggregation and applying it into existing multidimensional data. Those summarisations can be specified by a user and transformed in a form of a new dimension reusable with the already existing ones. Such application of OLAP exploration and multidimensional analysis conducts strength and power proved on industrial example (telecommunication company's customer buying habits).

Application of OLAP analysis into existing DBMS systems require deploying additional server, which generates additional analysis costs – data extraction time, maintenance costs. However, there is a way of adding OLAP analysis to existing database system without additional server deployment. Authors of [8] propose approach to perform OLAP exploration methods based on User Defined Functions (UDFs). In performance comparison between traditional approach and UDF-based one results presents that UDF require only one or two accesses to data comparing to access to data on each aggregation required by SQL-based approach.

3.2. Examples of OLAP Data Model

Researchers extend basic conceptual multidimensional model presented by [12] in search of bigger flexibility or better performance to satisfy bigger requirements of developing world of decision-making systems. In this section several examples of different approaches on OLAP data models will be described.

In the OLAP analysis systems multidimensional data is modelled based on current implementation OLAP server and data warehouse (relational – ROLAP or multidimensional – MOLAP). Such approach requires taking into the consideration physical organisation of data instead of considering only a logical layer. Authors of [6] propose a systemic method on describing logical part of data not bearing in mind how data is stored. Investigation starts with creation of E-R scheme describing a relational database. In that E-R diagram researchers try to discover all facts and dimensions possible, based on which the E-R scheme is restructured to present detected facts and dimension in more visible way.

In [34] proposal of data cube model and simple algebra supporting provided model is presented. Such approach is claimed to work as basis for future data cube implementations. In presented algebra researchers made assumed that all defined operators have to perform as good on data cubes as relational operators on relational structures. Proposed operators can be described as followed:

- Restriction – restraining one or more.
- Rename – change attribute name to prevent duplication.
- Cubic product – creation of link between two cubes.
- Metric projection – limitation of output data.
- Difference – detection of differences between two cubes.
- Aggregation
- Force – transformation of dimension into measures.
- Extract – transformation of measure to dimension.

Authors of presented claim that their algebra has the same level of expressivity as relational one and is closed (all operations on data cubes can be expressed using provided operators). Such approach provide simple algebra applicable as a foundation for future extension of presented research.

Conventional OLAP data model approach does not differentiate information about classifiers and characterizers. In [1] researchers propose object-oriented treatment of multidimensional cube. CubeStar in comparison to normal concept of OLAP data model points out drawback mentioned previous in paragraph. Basic structure in proposed concept is Multidimesional Object (MO), which also allow to perform OLAP operations changing property of object:

- Drill-down/Roll-up – change of granularity

- Split/Merge – add\remove property
- Slice/Unslice – change range of MO

To query object-oriented OLAP cube modified SQL-like language is introduced – Cube Query Language (CQL). CQL has many similarities but one of the most important differences is FROM clause – in CQL consist of dimensions covering current context of a query instead of additional relation information in normal SQL query. CubeStar does not store subsets of data (limited only by dimensions) but queries' answers. That Java implemented conceptual OLAP server results in much better performance than traditional OLAP server approach.

For database systems deployed before development of OLAP analysis it might be difficult to apply OLAP multidimensional model in relational database. Researchers in [29] propose an "SQLm" – Multidimensional SQL language, which provides compatibility with legacy databases. This high-level, user-oriented concept performs automatic aggregations and supports irregular dimension hierarchies. There three main definitions required to understand proposed model: summarizability (combined results of higher-level); aggregations; covering (no hierarchy path skips level); strict (no dimension values has more than one parent value from the same level). Generalized projection definition preventing unsafe aggregations is correct when inspected aggregation does not have missing or duplicated facts. That incidents might occur due to handling of irregular hierarchies.

Demands of analyst from OLAP analysis servers becomes more and more challenging – shorter query responses on volumes of data increasing promptly. Researchers in [37] propose ParaCube framework architecture applicable to many servers, which might be treated as one source of OLAP data source. ParaCube is based on parallel calculations on multiple servers (sibling servers) and grouping chunks of result into final query response. Sibling servers are servers warehousing whole replicated dimension table and part of fact table (local sibling cube). Those smaller subcubes are aimed to not have any common parts not to overlap query results. Such parallelism confirmed by experiment results presents better performance than big one machine calculation.

4. USING DATASTORY PATTERNS IN LOSD ANALYSIS

According to[26], data stories are artefacts for revealing and communicating insights gained from the analysis of data-sets obtained from the public domain, crowdsourcing or big data sources. From the same source and [17], data storytelling is a structured approach comprising elements including data, visuals and narratives for communicating insights obtained from data. Data stories could help to inform, explain, persuade or engage the target audience.

4.1. Linked Data Story Patterns

In our investigation about patterns in data storytelling, which could be applicable in exploring and analysing LOSD, a number of patterns were identified. We describe these patterns below grouped by their sources in literature.

Authors of [14] defined three types of data stories:

- Hero Stories are used to show that there is a solution to an existing problem. The "Hero" is the data and technology, which can solve the identified problem. The first key point of this type of story is to show the audience what is the actual problem. When the reader is aware of the problem, the effect of the "hero" (a specific solution or silver bullet) is then presented. An example of the problem or societal challenge could be climate change.
- Learning Stories involves making the audience aware of an existing problem. However, compared with Hero Stories, the presented crisis will not get solved by a "hero" but through participation.
- Horror Stories are used to exaggerate a societal or public problem for urgency. Authors employ this story type when they want to discuss very serious problem. In horror stories, no effect of intervention are presented.

Researchers in this paper [27] are using data stories as:

- Instructional Tool – this story type is aimed at creating stories as instructions to prevent bad decision/actions in the future. Information hidden in the data can show what mistakes were made in the past and by making stories authors show wider audience that there was a problem in the past and this is how you should avoid it. This could be in the choice of a specific policy choice in a particular context.
- Intervention – This story type aims to show the scale of a particular problem in the society, for instance a public health issue or disaster as a tool to convince vulnerable communities. Such stories encourage citizens to realise how big the problem is and that there is no ready solution to it. This pattern can be place between Learning and Horror Stories.

In [31], the authors discuss three general types of data stories:

- Martini Glass Structure – this story type starts from narrow author's narrative and later expands to allow reader to freely explore the data. Storyteller can only suggest which path is the best one.
- Interactive Slideshow - divide story in slides with author's narrative. Reader can move to another part (slide) of the story when he decides. This type is optimum for big sets of data or complex stories in order to avoid user getting lost.
- In Drill Down Story author declares possible interactions between user and the data. User, as in interactive slideshow, decides the pace of storytelling. Storyteller only present the main theme of the story and later reader takes over the control.

Also in [2], three ways of telling stories via created visual artefacts were presented:

- Author Driven – this type assumes no interaction with the audience. It consists of determined in advance order of narration and no "heavy message" is delivered.

- Reader Driven – this way of telling data stories enables interaction with the user. The user is able to relate emerging patterns in the interface to specific events. Audience is allowed to move freely around story and author can only point the suggested way of interpreting the data.
- Hybrid experience is a combination of two story patterns mentioned earlier. The author leads reader more than in Reader driven approach but reader has still some freedom to explore data himself.

There are two clear categories of story patterns from the reviewed literature above. The first set of data story patterns in [14] & [27] are named to reveal the nature or essence of the story. The nomenclature of second set of patterns in [31] & [2] are such that they reveal the nature of interactions afforded by the stories. They also suggest what the end-users can do with the data story artefact (visualisation or interactive table). In [26], a typology of data story telling patterns was described by the authors. The typology identified two categories of data story patterns including those characterised by on the nature of the data stories and the other based on the nature of analysis to be accomplished.

4.2. Categorization of discovered patterns

The patterns defined in 4.1 are somewhat high-level. Authors of [26] created typology, based on [10] and [11] which are used for the exploring LOSD. The Table below present some of the applicable "analytical" story patterns in Linked Data approach and gives an example or real-life condition applicable to this pattern.

Patterns categorized in Table 1 provide a foundation for domain-specific language [21] reducing the need for technical knowledge to understand and generate insight from LOSD of interest. Having a pattern language for data storytelling will allow for reusability of patterns as modules to compose more complex stories based on simpler story patterns defined below. Complex analysis can be expressed as sequence of simple data story patterns making it understandable for the larger audience without specialist knowledge.

Implementation of design patterns is valuable to many domains creating cross-disciplinary language describing complex concepts without deep knowledge about domain.

The identified data story patterns could be extended or modified to satisfy targeted data portal needs.

In the next section, we provide an example that demonstrates compositional aspect of story patterns allowing creating complex pattern pipelines.

5. CASE STUDY

We describe in this section how some of the data story patterns described in Table 1 could be applied in exploring and analysing three related linked open statistical data (or data cubes) published by the Irish Central Statistics Office. Information included in data might be valuable for citizens investigating situation in their neighborhood or organizations creating new policies. Presented data consist information about:

- Highest educational level completed,
- Principal economic status,
- Employment level in each industry.

Other dimensions are identical across Cubes–geographical areas and gender. If these three datasets were not published as RDF Cube, relating between them would be difficult and some information might be missed. Interlinking allow end-user to access interesting across three datasets.

Example queries are presented below:

- Relation between educational level and industry employment per each county.
  Local authorities might be interested in investigating the link between education and employment in their own county to better understand situation and find a way to improve it (support policy making process). If there is a need to investigate deeper the problem (start big and drill down), data can be organized in smaller areas to better understand which parts of county should be the main objective of new campaign.
- Industry employment and principal economic status
  Organisations handling social services might be interested in analysing which areas need most (start from areas with most needs and go to areas with the least needs) support to attract companies creating new job offers and reducing unemployment levels.
- Industry employment categorized by sex
  Women rights institutions might prepare juxtaposition which positions are dominated by men and encourage women to apply for this position to level up differences in employment levels.

There are multiple methods to achieve required output for example cases. Story patterns can be defined as query templates on the data portals which users may select to execute. Example story patterns are presented below as a sequence:

1) GET League Table (get top5 counties)
2) Narrating change over time (evaluate level changes in time)
3) Start big and drill down (investigate reasons for discovered issues)

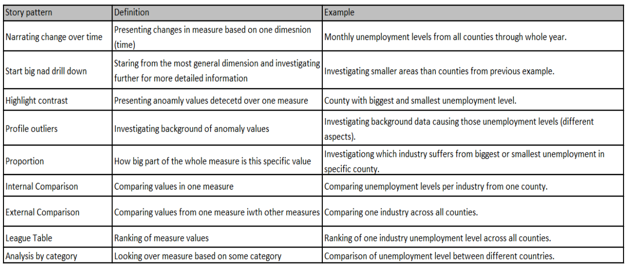
Example pipeline based on sample data is presented below in Figure 2. All county URIs are prefixed with: "http://data.cso.ie/census-2011/classification/areas/CTY". They have been omitted in the tables for brevity. These county URIs allows for the integration across the three data cubes.

The application of the first story pattern returns the top 5 countries in terms of gender disparity in the construction sector. The next story pattern returns the percentage change in the gender disparity from 2011 to 2012. The third and final pattern drills down into economic status and educational attainment across gender to better understand link between education and availability of jobs.

The implementation of these pattern does not require any technical knowledge. The user simply has to select parameters for the patterns (e.g. which sector to examine or range of dates to consider for changes over time).

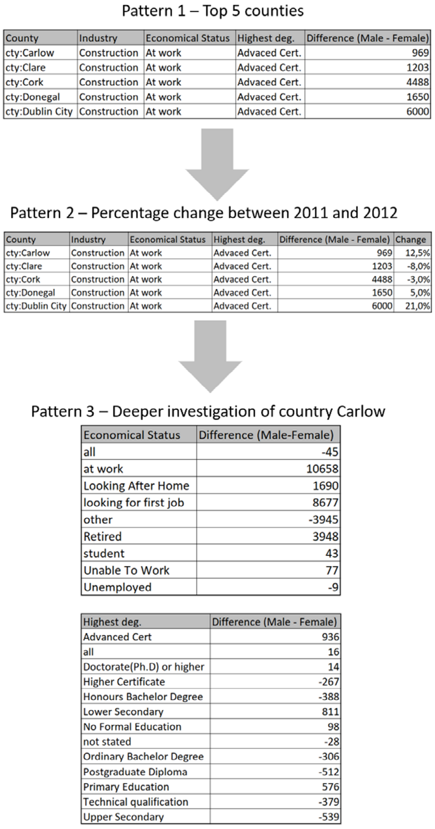
The pattern composition to create pipelines will be implemented based on rule engines and grammar [36]. Essentially, users will select storytelling pattern on the data platform which will be transformed or unfolded into a sequence of RDF C ube operations. Using data-pipeline architecture, output of one step becomes input for next pattern. In summary, the use of these patterns shields off the regular non-technical end-user from the technical details about graph databases, RDF operation or SPARQL queries. User is required only to have basic data literacy skills to understand output data [36].

Figure 2 provides a schematic of how data story pattern can be implemented on open data portals.



Story pattern	Definition	Example
Narrating change over time	Presenting changes in measure based on one dimesnion (time)	Monthly unemployment levels from all counties through whole year.
Start big nad drill down	Staring from the most general dimension and investigating further for more detailed information	Investigating smaller areas than counties from previous example.
Highlight contrast	Presenting anoamly values detecetd over one measure	County with biggest and smallest unemployment level.
Profile outliers	Investigating background of anomaly values	Investigating background data causing those unemployment levels (different aspects).
Proportion	How big part of the whole measure is this specific value	Investigationg which industry suffers from biggest or smallest unemployment in specific county.
Internal Comparison	Comparing values in one measure	Comparing unemployment levels per industry from one county.
External Comparison	Comparing values from one measure iwth other measures	Comparing one industry across all counties.
League Table	Ranking of measure values	Ranking of one industry unemployment level across all counties.
Analysis by category	Looking over measure based on some category	Comparison of unemployment level between different countries.

Tab. 1: Data Story patterns presentation



Pattern 1 – Top 5 counties

County	Industry	Economical Status	Highest deg.	Difference (Male - Female)
cty:Carlow	Construction	At work	Advaced Cert.	969
cty:Clare	Construction	At work	Advaced Cert.	1203
cty:Cork	Construction	At work	Advaced Cert.	4488
cty:Donegal	Construction	At work	Advaced Cert.	1650
cty:Dublin City	Construction	At work	Advaced Cert.	6000

Pattern 2 – Percentage change between 2011 and 2012

County	Industry	Economical Status	Highest deg.	Difference (Male - Female)	Change
cty:Carlow	Construction	At work	Advaced Cert.	969	12,5%
cty:Clare	Construction	At work	Advaced Cert.	1203	-8,0%
cty:Cork	Construction	At work	Advaced Cert.	4488	-3,0%
cty:Donegal	Construction	At work	Advaced Cert.	1650	5,0%
cty:Dublin City	Construction	At work	Advaced Cert.	6000	21,0%

Pattern 3 – Deeper investigation of country Carlow

Economical Status	Difference (Male-Female)
all	-45
at work	10658
Looking After Home	1690
looking for first job	8677
other	-3945
Retired	3948
student	43
Unable To Work	77
Unemployed	-9

Highest deg.	Difference (Male - Female)
Advanced Cert	936
all	16
Doctorate(Ph.D) or higher	14
Higher Certificate	-267
Honours Bachelor Degree	-388
Lower Secondary	811
No Formal Education	98
not stated	-28
Ordinary Bachelor Degree	-306
Postgraduate Diploma	-512
Primary Education	576
Technical qualification	-379
Upper Secondary	-539

Figure 1: Example Story Pattern Pipeline

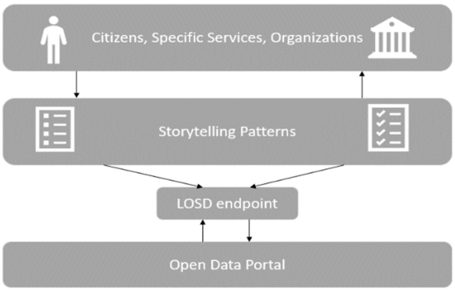


Figure 2: Proposed framework visualization

6. Discussion

The main motivation for our work on data storytelling patterns and their implementation on open data portals is to simplify or mediate access and analysis of datasets published on these portals. The focus on linked open statistical data is based on the wide availability of these categories of dataset, their high public value and their relative high quality when compared with other datasets. While raw and fine-grained dataset are useful for developers and producers of data products and services, the aggregated or summary forms statistical datasets make them simpler for consumption and more valuable for the non-technical end-users.

As reported in Section 2, there are some ongoing efforts at simplifying the consumption of linked open statistical data or data cubes. For instance in [32], a set of tools including cube browsers and explorers were presented. Unfortunately, the use of these tools require some level of technical data and analytical skills and competences that are rare to find in ordinary end-users.

In mapping existing data story patterns, we have identified at three different typologies of data story patterns in literature. These typologies are based on the nature, exploration and type of analysis to be carried. In our view these typologies are complementary and can be used as a design framework for data story pattern interface on the open data portals. For instance analytical patterns like those shown in Section 5 could support patterns considered as exploratory or those classified under the nature-oriented patterns. Figure shows this relationship

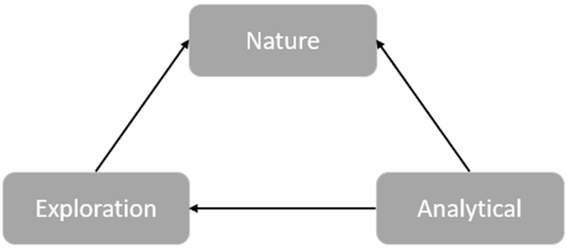


Figure 3: Categories of Data Story Patterns

We expect current and next generation data portals to support data storytelling features in the form described in our paper. The implementation of these data story telling features as plug-in to portals with the possibility to reuse patterns across data portals templates.

We believe the use of data story telling patterns in consuming open (statistical) data will empower citizens, institutions and other entities to use available high quality statistical data resources control, verify, investigate, support, understand, identify problems and issues in local environment.

7. CONCLUSION

We have shown how data stories could serve be used to simplify analysis and exploration of linked open data. In our opinion, providing data story patterns for use on open data portals should increase the adoption and use of open data.

This work does not only contribute concrete ideas for improving open data use but also contributes to data story telling literature. We envisage further research in the area of realizing the implementation of some of the patterns described in this paper on major data platforms such as CKAN or Datasoft. We also see research emerging in the area of data story pattern composition.

ACKNOWLEDGEMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

REFERENCES

[1] Albrecht, J. et al. 1999. Management of multidimensional aggregates for efficient online analytical processing. Proceedings. IDEAS'99. International Database Engineering and Applications Symposium (Cat. No.PR00265). (1999), 156–164. DOI:https://doi.org/10.1109/IDEAS.1999.787264.

[2] Balduini, M. et al. 2015. CitySensing: Fusing City Data for Visual Storytelling. IEEE Multimedia. 22, 3 (2015), 44–53. DOI:https://doi.org/10.1109/MMUL.2015.54.

[3] Bizer, C. 2009. The emerging web of linked data. IEEE Intelligent Systems. 24, 5 (2009), 87–92. DOI:https://doi.org/10.1109/MIS.2009.102.

[4] Brolcháin, N.Ó. et al. 2017. Extending Open Data Platforms with Storytelling Features. Proceedings of the 18th Annual International Conference on Digital Government Research - dg.o '17. (2017), 48–53. DOI:https://doi.org/10.1145/3085228.3085283.

[5] Brunetti, J. et al. 2012. The Linked Data Visualization Model. International Semantic Web …. (2012).

[6] Cabibbo, L. et al. 1998. A Logical Approac h to Multidimensional Databases. Informatica. (1998), 183. DOI:https://doi.org/10.1007/BFb0100972.

[7] Chaudhuri, S. and Dayal, U. 1997. An overview of data warehousing and OLAP technology. ACM SIGMOD Record. 26, 1 (Mar. 1997), 65–74. DOI:https://doi.org/10.1145/248603.248616.

[8] Chen, Z. et al. 2009. Fast and dynamic OLAP exploration using UDFs. Proceedings of the 35th SIGMOD international conference on Management of data - SIGMOD '09. (2009), 1087. DOI:https://doi.org/10.1145/1559845.1559989.

[9] E.F. Codd et al. 1993. Providing OLAP to User-Analysts: An IT Mandate.

[10] Exploring the 7 Different Types of Data Stories: http://mediashift.org/2015/06/exploring-the-7-different-types-of-data-stories/. Accessed: 2018-11-03.

[11] Gray, J. et al. The data journalism handbook : [how journalists can use data to improve news].

[12] Gyssens, M. 1997. A Foundation for Multi-Dimensional Databases. (1997), 106–115.

[13] Jain, P. et al. 2010. Linked Data is Merely More Data. Linked Data Meets Artificial Intelligence. Technical Report SS-10-07, AAAI Press. (2010), 82–86.

[14] Janda, K.B. and Topouzi, M. 2015. Telling tales: Using stories to remake energy policy. Building Research and Information. 43, 4 (2015), 516–533. DOI:https://doi.org/10.1080/09613218.2015.1020217.

[15] Kalampokis, E. and Tarabanis, K. 2017. Visualizing Linked Open Statistical Data to Support Public Administration. (2017).

[16] Kim, H. et al. 2016. Towards an enterprise entity hub: Integration of general and enterprise knowledge.

[17] Lee, B. et al. 2015. More Than Telling a Story: Transforming Data into Visually Shared Stories. IEEE Computer Graphics and Applications. 35, 5 (Sep. 2015), 84–90. DOI:https://doi.org/10.1109/MCG.2015.99.

[18] Leonhardi, B. et al. 2010. Augmenting OLAP exploration with dynamic advanced analytics. Proceedings of the 13th International Conference on Extending Database Technology - EDBT '10. (2010), 687. DOI:https://doi.org/10.1145/1739041.1739127.

[19] Linked Data - Design Issues: 2006. https://www.w3.org/DesignIssues/LinkedData.html. Accessed: 2018-07-02.

[20] Lomet, D. et al. 2013. Data Engineering. Society. 36, 4 (2013).

[21] MARJAN MERNIK, JAN HEERING, A.M.S. 2007. When and How to Develop Domain-Specific Languages. 37, 4 (2007), 5. DOI:https://doi.org/10.1145/1118890.1118892.

[22] Newell, B. and Siri, J. 2016. A role for low-order system dynamics models in urban health policy making. Environment International. 95, (2016), 93–97. DOI:https://doi.org/10.1016/j.envint.2016.08.003.

[23] Ni, L. et al. 2013. Visualizing linked data with JavaScript. Proceedings - 2013 10th Web Information System and Application Conference, WISA 2013. (2013), 211–216. DOI:https://doi.org/10.1109/WISA.2013.48.

[24] Ojo, A. et al. 2016. Pathologies of Open Data Platforms and Desired Transparency-Related Affordances for Future Platforms. Proceedings of the 17th International Digital Government Research Conference on Digital Government Research (New York, NY, USA, 2016), 538–539.

[25] Ojo, A. et al. 2016. Realizing the Innovation Potentials from Open Data: Stakeholders' Perspectives on the Desired Affordances of Open Data Environment. Collaboration in a Hyperconnected World: 17th IFIP WG 5.5 Working Conference on Virtual Enterprises, PRO-VE 2016, Porto, Portugal, October 3-5, 2016, Proceedings. H. Afsarmanesh et al., eds. Springer International Publishing. 48–59.

[26] Ojo, A. and Heravi, B. 2017. Patterns in Award Winning Data Storytelling: Story Types, Enabling Tools and Competences. Digital Journalism. 0811, (2017), 1–26. DOI:https://doi.org/10.1080/21670811.2017.1403291.

[27] Palacios, J.F. et al. 2015. Storytelling: A Qualitative Tool to Promote Health Among Vulnerable Populations. Journal of Transcultural Nursing. 26, 4 (2015), 346–353. DOI:https://doi.org/10.1177/1043659614524253.

[28] Paulheim, H. 2013. Exploiting Linked Open Data as Background Knowledge in Data Mining. CEUR workshop proceedings DMoLD 2013 : Proceedings of the International Workshop on Data Mining on Linked Data, with Linked Data Mining Challenge collocated with ECMLPKDD 2013. (2013), 1–10.

[29] Pedersen, D. et al. 2002. A Powerful and SQL-Compatible Data Model and Query Language for OLAP. Thirteenth Australasian Database Conference (ADC2002). 5, November 2015 (2002), 121–130. DOI:https://doi.org/10.1145/563932.563920.

[30] Resource Description Framework (RDF): Concepts and Abstract Syntax: 2004. https://www.w3.org/TR/rdf-concepts/. Accessed: 2018-07-25.

[31] Segel, E. and Heer, J. 2010. Narrative visualization: Telling stories with data. IEEE Transactions on Visualization and Computer Graphics. 16, 6 (2010), 1139–1148. DOI:https://doi.org/10.1109/TVCG.2010.179.

[32] Stasiewicz, A. et al. 2018. Using Linked Statistical Data to Improve Marine Search and Rescue Operations in Ireland. Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance - ICEGOV '18. (2018), 412–418. DOI:https://doi.org/10.1145/3209415.3209511.

[33] The RDF Data Cube Vocabulary: 2014. https://www.w3.org/TR/vocab-data-cube/. Accessed: 2018-08-03.

[34] Thomas, H. et al. 2001. Processing in Decision Support Databases A Conceptual Model and Algebra for On-Line Analytical Processing in Decision Support Databases. July 2015 (2001).

[35] Trinh, T.D. et al. 2013. Linked Widgets - An Approach to Exploit Open Government Data. Proceedings of the 15th International Conference on Information Integration and Web-based Applications {&} Services (iiWAS2013). (2013), 438–442. DOI:https://doi.org/10.1145/2539150.2539252.

[36] Wickham, H. 2010. A Layered grammar of graphics. Journal of Computational and Graphical Statistics. 19, 1 (2010), 3–28. DOI:https://doi.org/10.1198/jcgs.2009.07098.

[37] Zhang, Y. et al. 2010. ParaCube: A scalable OLAP model based on distributed aggregate computing with sibling cubes. Advances in Web Technologies and Applications - Proceedings of the 12th Asia-Pacific Web Conference, APWeb 2010. (2010), 323–329. DOI:https://doi.org/10.1109/APWeb.2010.31.