2022 IEEE/ACM 44th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS) | 978-1-6654-9594-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICSE-SEIS55304.2022.9793895

# Open Data Inclusion through Narrative Approaches

Annika Wolff
Natasha Tylosky
Tanvir Hasan
annika.wolff@lut.fi
natasha.tylosky@lut.fi
md.tanvir.hasan@lut.fi
LUT University
Lappeenranta, Finland

## ABSTRACT

Open data is published with the intention that it can be used by everyone. In reality, various barriers exclude some people from its use. This short paper will examine common reasons why people struggle to make use of open data and propose a conceptual framework to guide narrative-driven curation of open data sets to overcome these barriers. These principles are currently being used to guide the creation of new interfaces to open data that can be used by people with differing levels of data literacy, thereby increasing inclusion of open data sets.

## CCS CONCEPTS

• **Human-centred computing**; • **Human computer interaction (HCI)**;

## KEYWORDS

datasets, open data, narrative, search

**ACM Reference Format:**
Annika Wolff, Natasha Tylosky, and Tanvir Hasan. 2022. Open Data Inclusion through Narrative Approaches. In *Software Engineering in Society (ICSE-SEIS'22), May 21–29, 2022, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3510458.3513022

## 1 LAY ABSTRACT

Within the last decade or so, there has been a growing trend for governments and scientists to open up their data so that they can be used by anyone. This might include data about schools, local population statistics, traffic accidents, air pollution and so forth. The benefit is that people can check information from the source, rather than relying on secondary interpretations that may be biased or wrong. However, open data is not as inclusive as it could be. One type of exclusion is lack of representation of some people within a data set, limiting how it can be used. Another type of exclusion is related to lack of skills for making use of data, which could affect anyone but especially those with less formal education. This paper proposes how to make such data easier to use by anyone, using approaches that humans use every day, namely narrative and stories.

## 2 INTRODUCTION

The term *open data* generally denotes data in an open format that can be freely used, re-used and shared, by anyone and for any purpose. Some commonly promoted benefits of open data include improving government transparency [10], and driving the innovation of new smart apps and services, especially in solving local problems [15]. Open data typically covers a wide range of topics, including environment, economy, transport, statistics and health. It is often made available through open data portals, which may be specialised, for example as *open government data* or *open research data*. What is common amongst these examples is that access and use of the data should be *equitable* and *inclusive*. In other words, anyone can use it no matter who they are.

Despite such aims it has been found that whilst open data provision may be equal - in the sense that everyone has a similar opportunity to get hold of it - it is far from equitable. This is because people have different needs in their use of data and these are not always met. One reason for this is that in the development of open data portals "the user's view is largely ignored" [3]. Thus, whilst open data is often produced and stored with the expectation that people will be able to find and reuse it, this is often not the case. This paper will explore some of the common factors that impact data reuse and propose an approach based on narrative-principles to make open data easier to find and use.

## 3 COMMON BARRIERS AGAINST USING OPEN DATA

This section explores some of the barriers to open data use identified from key literature in the field.

**Barrier 1 - people have different technical skills for using data.** The public encounter many barriers in obtaining and interpreting data. Even when it is easily available it often requires specialist tools and knowledge for making sense of it, which puts it out of reach of many [3]. Some data, especially that which is large and complex, requires the use of statistical techniques, or knowledge of how to combine different data sets to create additional value from it [4]. As a result, only those with such technical knowledge can process open data can make sense of it, and many people have not had the opportunities for acquiring the necessary data literacy skills. The time and resources needed to get started

*ICSE-SEIS'22, May 21–29, 2022, Pittsburgh, PA, USA*
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9227-3/22/05…$15.00
https://doi.org/10.1145/3510458.3513022

in using some open data sets may therefore be prohibitive. The fact that not everyone has the technical skills or access to tools to work with data greatly affects its inclusivity and potential for reuse. *People need access to better learning resources, but at the same time - where possible - it is necessary to also reduce the need for specialist skills and tools.*

**Barrier 2 - open data can be hard to find, even when looking in the right place.** Most open data portals work as a kind of archive in which the user can search for the data sets they need using search, filters or categories. Since the search for data sets is the main task of the user on the site, the search should be fast and convenient. But many open data portals are better designed for the task of conveniently storing data and with less emphasis on the service provided to users when they want to search for and use the data [8]. A typical problem encountered is that search sometimes only works on the title, and the title itself is often not accurate. This confuses the user, especially the inexperienced. Also, while there are some standards for open data storage, the interfaces are far from uniform and even the standard ones that do exist are 'not inspiring' [7]. Therefore, even if you have learned to use one, then this knowledge does not necessarily translate to the next. *To solve this problem, data should be better described and the search process should be more intuitive and uniform across different sites.*

**Barrier 3 - open data has hidden imperfections.** According to Feinberg [2] data may be biased from the moment of its collection. Choices are made about when, where, how and from whom data will be obtained. This may place limitations on its potential reuse, or worse its use can even amplify a bias inherent in the data [9]. Similarly, if people are monitoring a species (as happens a lot in citizen science) then the data set will reflect where people have looked and found a 'positive' occurrence, rather than the places where they did not look, or did not find anything. Such gaps are easier to understand in the original data collection context but important information may be lost when storing the data for reuse. Similarly, sampling and other selection errors during this stage may lead to unrepresentative data. *To overcome this problem, it is important that as much original context as possible is captured and made available alongside stored data to help those that did not participate to its collection.*

**Barrier 4 - it can be difficult to judge relevancy of a data set returned by search.** When a user submits a query to an open data portal, it may be hard to judge whether the returned data sets are a good match. The usefulness of information depends on a range of factors, including how the interface of the portal is designed to show the metadata of the data set, the quality of this metadata [13] and the richness of any free text description used to describe the data. Most data owners tend to take the path of least resistance and publish data in its original format, ignoring the opportunity to make data available in more readable formats. Therefore, in order to explore data further, it is often necessary to first download the data and then to visualise it with separate tools as most open data portals don't allow to visualise data directly from the portal [6]. This can be time consuming even for those with good knowledge and tools and prohibitive if in the end many data sets are not of good relevance for the query. *To overcome this, there needs to be improved query loops, teasers and ideally data exploration tools embedded into open data portals. Also, as well as capturing data about context as described above, this information needs to be made easily available as part of search results as this will help the user to assess quality of the data.*

**Barrier 5 - open data portals do not support data discovery.** Open data portals are typically designed to support the type of search where the user knows what they are looking for. However, understanding how data informs and can be integrated into understanding problems and creating solutions may require more scaffolding for less experienced users and also a broader idea of what is happening when people are searching for and using open data. If open data is intended to increase participation, then open data practices need to be understood within a broader collaborative and problem-solving context and not just in terms of search and retrieval actions [11]. *Therefore, open data portals may benefit from recommender tools that suggest useful data sets, for example related to the data they are looking at but not necessarily directly related to their search - such as data related to the local community context in which the data is being explored.*

## 4 NARRATIVE-DRIVEN DATA CURATION

Given these barriers, it is clear that users of open data portals need to be better supported in evaluating data sets according to their relevance, usability and overall quality. In the rest of this paper we propose a narrative-driven data curation process aimed to overcome barriers against open data reuse. This adapts an existing approach that was previously developed and tested on cultural data sets (data about museum objects) [14]. In this paper we explore how a similar approach might be applied to more general data types.

The term *curation* originated as part of museum practice. It refers to activities related to i) assembling objects into a museum collection ii) managing the objects by researching, describing and preserving them and iii) presenting them within exhibitions for others to interpret. Museum curation requires taking objects out of the original context for which they were created and placing them alongside new objects to tell different stories with them. Similarly, when applied to data, the term curation tends to mean the practices of collecting and managing data so that it can be re-used and made sense of in alternative contexts. This step of curating data is critical in helping people to reuse the data and so it is important that it is done well.

This paper proposes that *narrative principles* can be applied to different data curation activities. People are inherently narrative thinkers and learn to make sense of the world through stories from an early age [12]. Narratives have certain common elements which include the *setting* (when and where the story takes place), *characters* (human or non-human) and *themes*. These are the elements of a narrative that provide coherence, so that the *events* that are happening in any given story may move around in time or place, follow different characters, or even explore different themes as long as most other aspects remain constant.

Figure 1 shows a process through which these narrative concepts of *setting* (time and place of data collection), *character* (primary attributes that characterise the data set) and *theme* (typical organisational themes, such as environment, transport, health) might be used as metadata to describe raw *data events* which in turn could be used to recommend *narrative groupings* of data sets, in other words

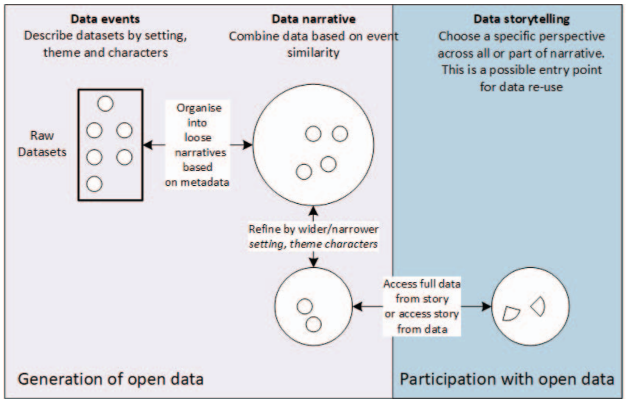

Data events
Describe datasets by setting, theme and characters
Data narrative
Combine data based on event similarity
Data storytelling
Choose a specific perspective across all or part of narrative. This is a possible entry point for data re-use
Raw Datasets
Organise into loose narratives based on metadata
Refine by wider/narrower setting, theme characters
Access full data from story or access story from data
Generation of open data
Participation with open data

**Figure 1: Conceptual framework of narrative-driven data curation**

data that could belong in a common story - or to filter data sets according to a narratively structured query. The narrative scope can be widened or narrowed along these same dimensions. Finally, when a story is told from data, all or part of a data set can be said to belong to that story. Thus, data stories can themselves provide a starting point for data exploration, with the data sets associated with them being turned into a narratively structured query, allowing someone engaging with the story to explore beyond what is being told. Through these mechanisms the narrative approach can support data sensemaking and telling stories from open data in relation to specific issues of concern, as long as such issues can be framed in terms of where, when, who and what they are about. The approach is designed to support overcoming some of the barriers described in the previous section. These concepts are now described in more detail.

**Data events**. Data events may describe the data itself in terms of setting (time and place), theme, and characters. In fact, both setting and theme are part of common open metadata standards, such as DCAT. However, raw data collection takes place within specific settings [5], which reflect that data is created by people, or their machines, in specific time, place and purpose. Additional 'data collection' events could therefore be associated to data that formally capture this information and which can be used as part of querying and search. This might highlight the interpretation or bias that has already been placed on the data through choices made about the purpose and manner of its collection. This is important to improve the re-usability of the data. Similarly, *data re-use events* themselves could capture reuse *settings* (where, when, by whom, why was data re-used). To further support data re-use, data teasers may be created that present this metadata as a 'data short story' to help judge quality, usability and relevance at a glance. A teaser in this case is a carefully selected snapshot or view of data that supplies key information. A teaser can be generic or it can be adapted to a specific context, for example tailored to a search query, to highlight the important aspect of data that has matched. **(Barriers 2-4).**

**Data narratives**. A number of data events may be related and organised into a loose narrative, based on overlap of narrative properties. These represent data sets that may be useful if combined, although the choice of combination depends on the exact problem being solved. Such data narratives could be identified through recommenders that use spatial, temporal and semantic proximity based on the metadata provided in the data events to find data close in settings, themes and characters. When combined with natural language search capabilities, such an approach could support data discoverability as well as make it easier for less technical users to construct complex searches covering many parameters in a more intuitive way For example, the query. 'What data do you have about bat species in Helsinki during the last ten years?' could be deconstructed to narrative principles and used to find the closest matches that cover the majority of concepts. Filters may widen or narrow the scope of a narrative across these different dimensions (e.g. expanding or narrowing the time period or themes). This could be used to support collaborative problem solving by framing the search in the context of the scope of the problem being solved. **(Barriers 1 and 5)**.

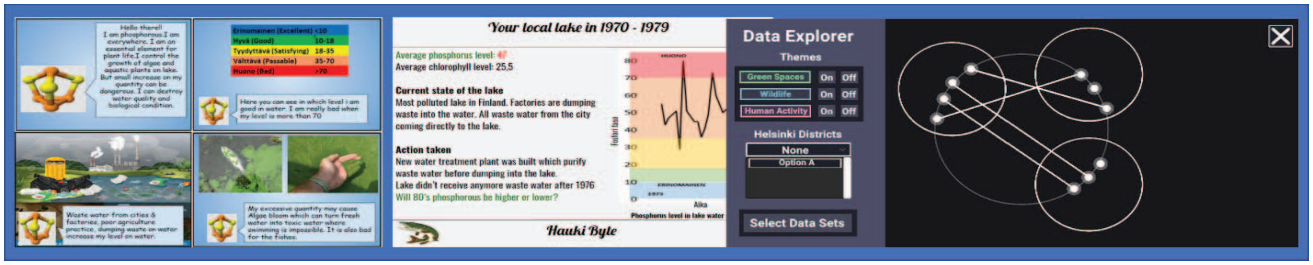
**Data story**. A data story is defined by a user and it is a specific interpretation of data that bounds a proposed narrative by determining the setting, theme and characters that are relevant to that particular viewpoint across the dataset(s). A data story is thus linked to the data sets and specifically the parts of the data that are used in its creation. A data story can *itself* act as an entry point for exploring data for re-use, which may be especially useful for less confident users who might struggle to start finding stories from raw data. By applying narrative principles in reverse, the user may start to expand the story in directions they find interesting or to discover new data sets that offer new interpretations across the data, supported again by a data recommender. Thus, each time data is explored and used in a story there are more elaborations and ideas on how different data may fit together and this in itself can support data discovery and data search. For any data that is found in an open data portal, it may be possible to see the one or more stories associated to it. **(Barriers 1,2, 4 and 5)**.

Figure 2 on the left-hand side shows an example of a data story created using a data comic approach, for exploring water quality data. It this has been realised as a physical activity, in which readers follow the characters of phosphorous and chlorophyll and their changes over 5 decades. This story can be explored further by bringing in different 'characters' e.g. lake elements or by moving to a new place (the lake was monitored in different places, and many lakes have similar monitoring done for them). On the right-hand side is an example of a data explorer that can be used to annotate and explore data sets using narrative principles.

## 5 DESCRIBING DATA

In order to achieve the above it is necessary to specify more closely how data will be described and how this information will be used in queries. Whilst it is beyond the scope of this paper to go into all the technical details the following offers some insights into the possibilities into five key activities:

(1) **Describe:** A data set is described by 4 elements of time, place, characters and themes. Key questions relate to how to level of granularity of description for all of these elements, but especially time and place. Temporal descriptions need to include when data collection started, when it ended, or if it



Hello there!! I am phosphorous.I am everywhere. I am an essential element for plant life.I control the growth of algae and aquatic plants on lake. But small increase on my quantity can be dangerous. I can destroy water quality and biological condition.

| | |
|---|---|
| Erinomainen (Excellent) | <10 |
| Hyvä (Good) | 10-18 |
| Tyydyttävä (Satisfying) | 18-35 |
| Välttävä (Passable) | 35-70 |
| Huono (Bad) | >70 |

Here you can see in which level i am good in water. I am really bad when my level is more than 70

Waste water from cities & factories, poor agriculture practice, dumping waste on water increase my level on water.

My excessive quantity may cause Algae bloom which can turn fresh water into toxic water where swimming is impossible. It is also bad for the fishes.

Your local lake in 1970 - 1979

Average phosphorus level: 47
Average chlorophyll level: 25,5

Current state of the lake
Most polluted lake in Finland. Factories are dumping waste into the water. All waste water from the city coming directly to the lake.

Action taken
New water treatment plant was built which purify waste water before dumping into the lake.
Lake didn't receive anymore waste water after 1976
Will 80's phosphorous be higher or lower?

Hauki Byle

Figure 2: Example of data story and data explorer for narrative driven exploration of data

is ongoing and also how frequently data is collected - is it one time, or periodically and if so what is the update period. Geographical descriptions may identify many very specific locations (e.g. locations of sensors from which data is collected periodically, specific buildings, landmarks or trees) or may refer broadly to regions without pinpointing any specific entity that the data was derived from (e.g. demographic data that is published on a regional level, such as population density, or number of people in energy poverty). Most metadata schemes define approaches that can be used.

(2) **Merge**: this refers to the act of merging narrative descriptions of two or more data sets to create a single query. This could be two or more data sets that belong to a story, or that have been returned by a search query and somehow validated by a user as being potentially useful.

(3) **Query**: a search query is comprised of four narrative aspects, which are *time*, *place*, *characters* and *themes*. A user might construct a query directly, for example using keywords, filters, or even natural language They could even use an existing data story as as a starting point for a new search, if they are looking to expand on that story. In this case, the content of the story could act in a similar way to a natural language query, or if the story is directly linked to narratively described data then those individual or merged data set descriptions can themselves be a query. Similarly, a recommender could construct a query in the process of finding data sets to recommend. There are NLP methods for extracting entities and time periods from text and categorising them and this same approach could be used to broaden the search beyond just data sets but to provide additional background information for extra context in understanding the returned data, something that has been suggested could be useful [7].

(4) **Search**: searching is the process of using the query to retrieve a set of search results. The search mechanism might use the percentage of overlap across all four elements, time, place, characters and theme and could be weighted in different ways, to provide more emphasis for example of thematic similarity and less on geographical.

(5) **Overlap**: describes the amount of narrative overlap between two data sets across the four elements of time, place, character and theme. There are different methods available for measuring spatial and temporal overlap of data and the most appropriate may be dependent on the context and the types of data available [1]. The descriptions of characters and theme can be compared using more simple measures such as cosine similarity of words.

(6) **Proximity**: in addition to identifying data that overlaps to a search query, it can be useful to identify and recommend data sets that are proximal to the query but not directly covered by it. For example, data that is from a neighbouring region or just outside the time period, but with no overlap. This supports the expansion of a data narrative, which could lead to new characters, areas, time periods being included to a new data story.

(7) **Recommend**: in addition to returning data that directly matches a query, a recommender may suggest data based on its proximity to a search result. In addition to proximity of settings, characters and themes, this could also include data that is proximal by prior use, for example data sets that have commonly been used together. This is where creating additional data events based on re-use, not just raw data collection, may be useful.

## 6 CONCLUSIONS AND FUTURE WORK

This approach of describing data according to narrative has the potential to help develop search interfaces that overcome barriers of using open data especially for less technical users. This will increase some aspects of data inclusion. Curation is an ongoing activity that should happen at different stages of the lifecycle, when it is collected, originally used and when it is re-used. The approach described in this paper supports story-driven querying and also story-based entry into data exploration, especially if supported by recommender systems. It has the potential to support all users of open data but especially those who have less technical expertise. The ideas presented in this paper will be formally tested in future work using a combination of controlled experiments and field studies.

## ACKNOWLEDGMENTS

This work has been funded by the Academy of Finland through the project ’Building the Civic Datascape’. This project has also received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 872500 within the project ParCos. We would also like to acknowledge the support and input of Helsinki Region Infoshare.

## REFERENCES

[1] Marcel Cardillo and Dan L. Warren. 2016. Analysing patterns of spatial and niche overlap among species at multiple resolutions. *Global Ecology and Biogeography* 25, 8 (2016), 951–963.

[2] Melanie Feinberg. 2017. A design perspective on data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2952–2963.

[3] Marijn Janssen, Yannis Charalabidis, and Anneke Zuiderwijk. 2012. Benefits, adoption barriers and myths of open data and open government. *Information systems management* 29, 4 (2012), 258–268.

[4] Jan Kučera. 2017. Analysis of barriers to publishing and re-use of open government data. *IDIMT 2017 Digit. Manag. Soc. Econ.-25th Interdiscip. Inf. Manag. Talks* (2017), 305–314.

[5] Yanni Alexander Loukissas. 2019. *All data are local: Thinking critically in a data-driven society.* MIT Press.

[6] Renata Máchová and Martin Lněnička. 2017. Evaluating the quality of open data portals on the national level. *Journal of theoretical and applied electronic commerce research* 12, 1 (2017), 21–41.

[7] Sebastian Neumaier, Vadim Savenkov, and Svitlana Vakulenko. 2017. Talking open data. In *European Semantic Web Conference*. Springer, 132–136.

[8] Anastasija Nikiforova. 2021. Smarter Open Government Data for Society 5.0: Are Your Open Data Smart Enough? *Sensors* 21, 15 (2021), 5204.

[9] Eirini Ntoutsi, Pavlos Fafalios, Ujwal Gadiraju, Vasileios Iosifidis, Wolfgang Nejdl, Maria-Esther Vidal, Salvatore Ruggieri, Franco Turini, Symeon Papadopoulos, Emmanouil Krasanakis, et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356.

[10] Kieron O'Hara. 2012. Transparency, open data and trust in government: shaping the infosphere. In *Proceedings of the 4th annual ACM web science conference*. 223–232.

[11] Erna Ruijer, Stephan Grimmelikhuijsen, Jochem Van Den Berg, and Albert Meijer. 2020. Open data work: understanding open data usage from a practice lens. *International Review of Administrative Sciences* 86, 1 (2020), 3–19.

[12] Roger C Schank. 1995. *Tell me a story: Narrative and intelligence.* Northwestern University Press.

[13] Thomas Schauppenlehner and Andreas Muhar. 2018. Theoretical availability versus practical accessibility: The critical role of metadata management in open data portals. *Sustainability* 10, 2 (2018), 545.

[14] Annika Wolff, Paul Mulholland, and Trevor Collins. 2013. Storyscope: using theme and setting to guide story enrichment from external data sources. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. 79–88.

[15] Anneke Zuiderwijk, Marijn Janssen, and Chris Davis. 2014. Innovation with open data: Essential elements of open data ecosystems. *Information polity* 19, 1, 2 (2014), 17–33.