Sustainability of (Open) Data Portal Infrastructures

Dataset Reuse: A Method for Transforming Principles into Practice

This study has been prepared by the University of Southampton as part of the European Data Portal. The European Data Portal is an initiative of the European Commission, implemented with the support of a consortium led by Capgemini Invent, including Intrasoft International, Fraunhofer Fokus, con.terra, Sogeti, 52North, Time.Lex, the Lisbon Council, and the University of Southampton. The Publications Office of the European Union is responsible for contract management of the European Data Portal.

For more information about this paper, please contact:

European Commission
Directorate General for Communications Networks, Content and Technology
Unit G.1 Data Policy and Innovation
Daniele Rizzi – Policy Officer
Email: daniele.rizzi@ec.europa.eu

European Data Portal
Gianfranco Cecconi, European Data Portal Lead
Email: gianfranco.cecconi@capgemini.com

Esther Huyer
Email: esther.huyer@capgemini.com

Written and reviewed by:
Laura Koesten
Email: l.m.koesten@soton.ac.uk

Elena Simperl
Email: elena.simperl@kcl.ac.uk

Johanna Walker
Email: j.c.walker@soton.ac.uk

Last update: 02.03.2020
www: https://europeandataportal.eu/
@: info@europeandataportal.eu

DISCLAIMER
By the European Commission, Directorate-General of Communications Networks, Content and Technology. The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use, which may be made of the information contained therein.

Luxembourg: Publications Office of the European Union, 2020

OA-02-20-168-EN-N ISBN: 978-92-78-42151-9 doi: 10.2830/407102

Note: this document is part of a series of research reports developed on the topic of "Sustainability of (open) data portal infrastructures", all of which are available on the European Data Portal at https://www.europeandataportal.eu/en/impact-studies/studies .

The series is made of the following reports:

1. A summary overview
2. Measuring use and impact of portals
3. Developing Microeconomic Indicators Through Open Data Reuse
4. Automated assessment of indicators and metrics
5. Assessment of Funding Options for Open Data Portal Infrastructures
6. Open data portal assessment using user-oriented metrics
7. Leveraging distributed version control systems to create alternative portals

Abstract

Portals have impact if the datasets they publish are used. In the first two reports, we have looked at ways to measure the economic impact of open data portals, arguing for the need to define granular indicators focusing on the extent to which datasets are reused. We also proposed a methodology to define microeconomic attributes and metrics for projects which use open data, which we have applied in scenarios in several EU countries. In this report, we consider reuse indicators and metrics which can be automatically assessed, and as such, as not bound to projects enabled by open data, but to the portals facilitating the use.

Automation is key to the ability to grow purposefully and adapt to user needs and feedback. It ensures that portal owners can routinely undertake assessments on large samples of datasets and activities and incorporate the findings into product roadmaps. We devise a methodology that assists portal owners with mapping high-level indicators from state-of-the-art data publishing literature and guides to lower-level observable features which a portal can keep track of and which co-relate with reuse. We show how the methodology could be applied to predict dataset reusability based on how they are published. By understanding which aspects of dataset publishing and use impact reusability, portal owners can improve their publishing practice, iterate over the design of their portals, and prioritise publishing and maintenance work.

Table of Contents

Abstract 3

1. Introduction 5

2. Method in a nutshell: from reuse principles to practice in 7 steps 7

3. Example 7

Step 1: Scope the assessment 7

Step 2: Define reuse metrics 8

Step 3: Collect reuse metrics 9

Step 4 – Define reuse indicators 9

Step 5: Cluster the datasets 12

Step 6: Analyse the data 13

Step 7: Derive recommendations 14

4. Conclusion 15

References 16

1. Introduction

An increasing amount of data is published openly on the web, ideally with the aim of reuse. One of the key challenges to its' uptake is supporting formats and capabilities to make it useful in as many contexts as possible (Shadbolt 2012). Reuse is more common in some domains than in others: Scientists reuse data of their peers to repeat previous experiments, propose new solutions, and derive fresh insights. Data is recognised as an asset in itself, cited and archived just like scientific literature. Developers define benchmarks and gold standards that everyone can use to establish to compare related approaches. They reuse such datasets to ensure that approaches remain comparable. Supervised machine learning, one of the most successful types of AI is dependent on the availability of relevant datasets to train algorithms. In this case, reuse is an economic necessity –deep learning architectures need to be pre-trained on large amounts of data and generating new datasets is too costly for most machine learning applications.

Reusability is stated as one of the four FAIR principles, a compilation of high-level best practices for making data *findable, accessible, interoperable, and reusable*. The "R" in FAIR gives guidelines on reusability include the following points, all focusing on metadata: (i) meta(data) are richly described with a plurality of accurate and relevant attributes, (ii) (meta)data are released with a clear and accessible data usage license, (iii) (meta)data are associated with detailed provenance, (iv) (meta)data meet domain-relevant community standards. The EDP in itself can be understood as a tool to improve the FAIRness of the over 1 million open government datasets it harvests.

While the FAIR metrics group[1] provides exemplary metrics for the FAIR principles, measuring FAIRness is not an established practice. There are also a variety of best practices and guidelines detailing data sharing and reuse principles, including the W3C best practices for data on the web or SharePSI or metadata standards for different purposes: general purpose standards such as Dublin Core[2] or DCAT[3], focusing on specific elements such as provenance (PROV[4]) or data quality[5] as well as domain specific extensions or standards.

Despite these efforts, portal owners and data publishers do not measure reuse routinely. Existing guidelines, indicators and metrics cannot be trivially mapped to observable features in the technical architecture of the publishing platform, which could be tracked and assessed automatically. Previous work [citeEDP1report] has suggested several solutions, including pixel tracking, dataset citations, and enforcing log-ins. These solutions have important limitations:

- Pixel tracking, and similar methods, operate at a granular level, and findings depend on the front-end design of the platform rather than on how useful the dataset is. More importantly, translating

[1] http://fairmetrics.org
[2] https://www.dublincore.org/groups/tools/
[3] https://www.w3.org/TR/vocab-dcat/
[4] https://www.w3.org/2001/sw/wiki/PROV
[5] https://www.w3.org/TR/vocab-dqv/

pixel-tracking insights into principles and practices to make datasets more reusable is hard, as the former is too low-level for the latter.

- Dataset citations, while an excellent idea, is not widespread outside scientific communities. While an incentives system for data citations is emerging in this space, it is unclear how it would transfer to open government data.
- The most used public sector datasets (such as urban transportation) often have excellent ecosystems that enable them to track usage in a less automated fashion (such as surveys, or app galleries). While the intense usage of their datasets, and the value of learning more about what features are most beneficial justify the cost of managing this tracking, this does not transfer to datasets that are less popular, as these cannot draw from a community of users for feedback. In the same time, the holders of these high-value data assets may not have the incentives to explore new tracking methods that would benefit other types of datasets.
- Finally, very few portals imply publish their own data – most provide a platform for data from a variety of sources, and some, such as the European Data Portal, are catalogues of datasets. Therefore, most portals are not in a position to implement tracking features such as log-ins.

Therefore, it is vital to address an alternative assessment approach, which focuses more on the reuse side of open data than the publishing, with automation support. This report presents such an approach. We introduce a method that helps a portal owner understand what makes a dataset more or less reusable, using engagement data they can track themselves. To apply the method, the portal needs to capture a minimum of engagement metrics, map higher-level dataset reuse indicators to such metrics and identify a subset that co-relate with reuse.

Automated assessment of reuse remains a substantial challenge. In an ideal world, a more end-to-end tracking of portal activities throughout the process would enable this. However, this requires new underlying structures, and while these may well be necessary eventually to ensure the sustainability of portals, the description of this goes beyond the remit of this report, which describes what can be achieved with the current technology, or with minimal adjustments. For these reasons, we have validated the method in a scenario which captures data about how people engage with datasets, for which such engagement data is easily available. We provide recommendations for portal owners to augment their publishing and portal design practice to support and enhance those features of a dataset that are quantifiably linked to higher engagement from users.

This report is organised as follows: we start by explaining how the method works in a nutshell. We then give an example for applying the method, which was informed by existing standards, best practices and guidance on how to make datasets easier to share and reuse and validated in a case study using datasets published and used (as well as engagement data) from GitHub. We show that it is possible to identify a basket of engagement metrics and predict the reusability of a dataset based on attributes such as: its structure, the way it was published, and its documentation.

2. Method in a nutshell: from reuse principles to practice in 7 steps

The method consists of the following steps, to be carried out by teams managing open data portals:

1. **Scope** the assessment exercise, for instance by deciding the specific collection of datasets that will be considered.
2. **Define reuse metrics**. These depend on the capabilities of your portal and the underlying technical infrastructure. If you cannot define direct metrics, think about proxy metrics. Run a study to validate them by exploring if they are quantifiably linked to reuse.
3. **Collect reuse metrics** (or proxies). For this, you need technical capabilities which may be built into the publishing software you're using, or aggregated metrics derived from lower-level system logs.
4. **Define reuse indicators**. These need to be measurable and will be used as features in the prediction model. In Section 3.1 we provide a list which can be used as a starting point, based on a comprehensive literature review.
5. **Analyse their distribution for the top-reused group of datasets**.
6. **Use a combination of those features to build a statistical model** to predict reusability.
7. **Derive recommendations** to datasets and publishing processes.

3. Example

Step 1: Scope the assessment

For exemplification purposes we chose an openly available corpus of datasets which have been shared via the GitHub platform. The corpus consists of 1.8 million data files, from over 87k repositories.[6] A repository may include one or more data files and is owned by a data publisher. A data publisher may create multiple repositories. All datasets were tabular data.

The corpus was created and pre-processed as follows: we used BigQuery to build an original list of non-forked[7] repositories that contain a CSV or XLSX or XLS file. We used the GitHub API to collect information about each repository in the original list. The resulting dataset consists of 87 936 repositories that contain at least a CSV, XLSX or XLS file, alongside with complementary information on their features (e.g. number of open and closed issues and license) from GitHub. We looked at those features as potential reuse indicators in step (5).

The resulting corpus contained more than two million data files. We then went ahead and excluded those with less than ten rows, after which we arrived at more than 1.8 million data files[8] (1373335 CSV files, 312870 XLSX files and 203865 XLS files). The percentage of data files in a data repository was on average

[6] A repository is a folder with multiple files in various formats.
[7] Forks are essentially copies of repositories. We looked at non-forked repositories to eliminate duplicates, which would have skewed the analysis.
[8] A valid excel file is one that has at least one sheet with 10 rows of data.

7.4% (med: 1.85%). This means that most dataset repositories had a number of other file types in the same repository, for instance containing code or documentation.

Step 2: Define reuse metrics

Most portals do not make any interaction data public. GitHub, as well as other data platforms which are about building a community just as much as they are about releasing datasets, provides a range of engagement metrics with each repository, which are indicative of usage.

In our example, the metrics describe different types of user activity that happens around a dataset:

- **number of forks** (copies of the dataset made by other users);
- **number of watchers** (i.e. subscribers) (users who have asked to receive notifications on a dataset);
- **number of stargazers** (users who have bookmarked a dataset); and
- **number of GitHub accounts that commit to the data repository** (users who have shared a version of the dataset in the repository).

The portal team would need to define their own reuse metrics or proxies, depending on their own goals and the capabilities of the portal infrastructure. For example, CKAN allows multiple users to update and refine a dataset if they are registered under the same institutional account, which means that engagement around a dataset can be tracked and documented. Other data platforms support features such as bookmarks (see Figure 1), followers, discussions etc.

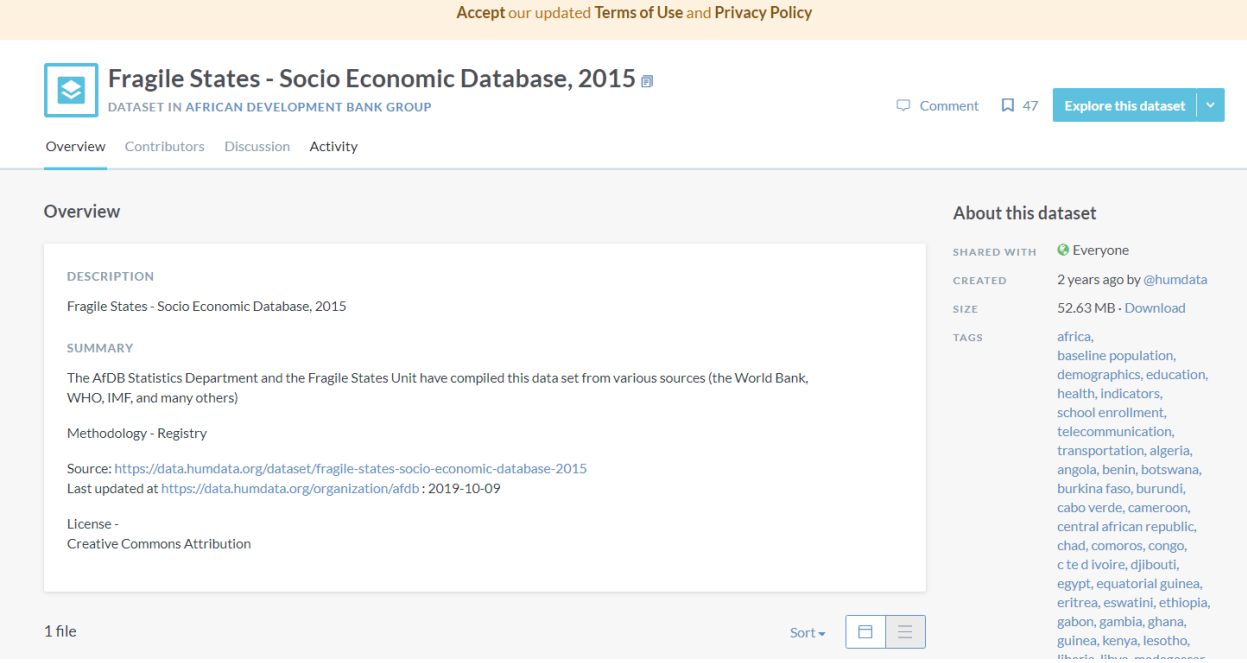

Figure 1: Example of a dataset shared via data.world. Note the discussion tab and the 47 people who bookmarked the dataset.

Step 3: Collect reuse metrics

For the four reuse proxy metrics identified in step (2) we collected the relevant data for all 1.8 million datasets in the 87k repositories:

- **number of forks** – we collected the forks count by calling the GitHub API iteratively.
- **number of watchers** - watching a repository registers the user to receive notifications on new discussions, as well as events in the user's activity feed and is called subscribers in the API[9]. We collected watcher count by calling the API iteratively.[10]
- **number of stargazers** - repository starring lets users bookmark repositories. Stars are shown next to repositories to show an approximate level of interest and have no effect on notifications or the activity feed.
- **number of GitHub accounts that commit to the data repository** - we counted the number of different email addresses which have committed on the master branch. We collected these counts by using regular expressions on each data repository .git file. Note that it is possible that the same person commits with different email addresses.

Step 4 – Define reuse indicators

We undertook an extensive literature review, which led to **39 indicators** grouped into **six themes** (see also Table 1):

[9] https://developer.github.com/v3/activity/watching/
[10] The GitHub API returns up to a 100 results per watchers or forks requests).

- Access;
- Documentation;
- Methodological choices;
- Versioning and provenance;
- Connections (links); and
- Others

Table 1: Reuse indicators. These are, across all different academic, white and green papers, as well as guides and technical standards, attributes of datasets and of the process through which these datasets came about, which experts recommend publishers to do to make their datasets more reusable. The assumption is that the better a dataset performs according to these indicators, the more reusable it will be.

REUSE INDICATOR	DESCRIPTION
Theme: Access	
License	i) available ii) allows reuse
Format / Machine readability	consistent format / single value type per column
Code available	for cleaning, analysis, visualisations
Unique identifier	for the dataset / IDs within the dataset
Download link / API	i) available ii) functioning
Theme: Documentation	
Description / ReadMe file	meaningful textual description (can also include text, code, images)
Purpose	purpose of data collection, context of creation
Summarising statistics	i) on dataset level, ii) on column level
Headers understandable	i) column level documentation (e.g. abbreviations explained) ii) variable types iii) how derived (e.g. codebook in social science
Missing values / null values	i) defined what they mean, ii) ratio of empty cells
Possible options / constraints on a variable	i) if data contains an "other" category
Geographical scope	i) defined, ii) level of granularity
Temporal scope	i) defined, ii) level of granularity
Time of data collection	when collected / what timespan?
Last update	information about data maintenance if applicable
Completeness of metadata	empty fields in the applied metadata structure?
Abbreviations / acronyms / codes	defined
Theme: Methodological choices	
Methodology	description of experimental setup (sampling, tools, etc) link to publication / project
Units and reference systems	i) defined, ii) consistently used
Representativeness / population	in relation to the total population / total population
Caveats	changes: classification/ seasonal or special event/ sample size / coverage /rounding

Cleaning / pre-processing	cleaning choices described, is the raw data available
Margin of error / reliability / quality control procedures	estimates vs actual measurements
Biases / limitations	different types of bias (i.e. sampling bias)
Data management	e.g. storage
Theme: Connections	
Relationships between variables defined	i) explained in documentation, ii) formulae
Cite sources	i) links or citation, ii) indication of link quality
Links to dataset being used elsewhere	e.g. in publications, community led projects
Contact	person or organisation, mode of contact specified
Theme: Versioning and provenance	
Version indicator	version or modification of dataset documented
Version history	
Prior reuse / advice on data reuse	(i) example projects (ii) access to discussions
Publisher / producer / repository	i) authoritativeness of source, ii) funding mechanisms / other interests that influenced data collection specified
Theme: Other	
Ethical considerations, personal data	i) data related to individually identifiable people, ii) if applicable, was consent given
Use of existing taxonomies/vocabularies	is this documented?
Quality metrics	i) consistent datatype per column, ii) amount of missing values, iii) check for outliers iv) confidence intervals
Visual representations	statistical properties of the dataset
Schema / Syntax / Data Model	defined
Duration of data storage	defined

We then mapped these indicators to how data is published, shared and used on GitHub. We narrowed down the list to retain only those indicators that can be observed and measured. The list contains 17 indicators, which we organised in three themes (see Table 2):

- Repository: properties of the folder in which the dataset sits.
- Documentation: in GitHub this is mostly in the form of a so-called ReadMe file.
- Datasets: this refers to the files in which the data was actually released.

Table 2: Observable reuse indicators for datasets published on GitHub

Theme	**Indicator on GitHub**
Repository	Age of repository
	Size of repository

	Licence
	Textual description
	Ratio of open to closed issues
	Ratio of data files to all files in a repository
	Aggregated size of all the data files in the repository
	Ratio of problematic files with respect to a particular
Documentation (ReadMe files)	Length of the documentation
	Unique URLs
	Language of the documentation (English or not)
	Number of coding blocks in a description (i.e. both inline and highlighting blocks)
	Number of images (i.e. Logo)
Datasets (data files)	Number of rows and columns of each individual data file
	Missing values
	Size of each data file
	Ratio of data files per repository

Just like in step (2), each portal will define their own reuse indicators, starting for orientation from Table 1 and adapting that list to relevant metrics, which are aligned with the goals of the project, the data publishing practice in the project, and the capabilities of the technical infrastructure. In [citereportask1] we also discuss other characteristics of useful metrics, which portal owners could take into account.

In our case study, we used indicators which are fairly generic, including size of the datasets, number of rows and columns, completeness, availability of documentation, licenses etc, which can be found, for instance, in CKAN metadata records as well.

Step 5: Cluster the datasets

We were able to group the repositories into 4 groups by level of user engagement with the repository.[11]

- Group 1 includes the repositories with the minimum of engagement with the repository.
- Group 2 included those with up to three counts in each reuse proxy metric.
- Group 3 those with up to nine counts in each category
- Group 4 includes all repositories with more counts.

This enabled us to define the highest ranking, most engaged with (and therefore, likely most used) repositories.

[11] Using an aggregated Borda count, which enables the aggregation of multiple ranked lists.

Step 6: Analyse the data

Looking at a statistical analysis of those repositories that are very likely to be reused showed a number of interesting results. For instance, the textual description of the data repository was longer, the repositories have a lower number of problematic files (meaning they can be opened with standard configurations), and the age of the repository does not correlate much with its reuse. There was also more "traffic" around the datasets visible, in terms of community engagement through opening and closing issues on the platform that notify others.

We defined each repository by vectors for each type of feature. Features correspond to the indicators from Table 2. The aggregated features for each of the three themes are summed up into a set using different neural network architectures to process them. The model then predicts which group a repository belongs to, based on this representation.

The model uses features from all three themes to learn what makes a dataset reusable in this particular context. For our GitHub analysis, the repository features (see Table 2) were found to be most predictive. The approach categorises a dataset repository into 1 out of 4 potential groups of reuse likelihood: Very likely to be reused; likely to be reused; moderately likely to be reused and unlikely to be reused.

We selected the 20 top ranked repositories according to our aggregated ranked list for a manual analysis of the ReadMe files to get a better understanding of those features that are not possible to assess automatically. We took the reuse indicators from Table 1 as primary categories to code for in the sample repositories.

For most features we used a feed forward neural net, only the description of the dataset repository (a short descriptive textual snippet usually provided by the repository owner) is treated differently using a neural network specifically for short text.

While this architecture needs to be tailored to a specific dataset repository and use case we present it as a prototype that is useful for other contexts from a modelling perspective. We combine a number of different feature types – counts, ratios, binary categories as well as short text snippets and tie them together to represent a dataset repository. Another added variable is the variation of tabular data files per repository, which reflects real world use cases.

We use those features in the model as these are provided by the GitHub API and tracked across this large number of dataset repositories we investigated. However, hypothetically many reuse indicators listed in Table 1 could be represented as part of this architecture if tracked across a large number of dataset repositories. This opens up a large space for both research in this area to develop the model further but also incentivises the tracking of reuse indicators to better understand their impact on real world reuse.

Combining all the available features (i.e. the general repository, data file and ReadMe features) enables our predictive model to achieve its highest accuracy score of just under 60%.

For larger repositories representing projects the ReadMe's included links to external documentation, such as a project website. We included the content of these resources in our analysis of documentation practices of they were easily accessible and pointed to in the ReadMe.

Step 7: Derive recommendations

While extensive documentation of all indicators mentioned in Table 1 would be ideal we propose to prioritise the following based on our analysis of the GitHub dataset repository corpus:

- A short textual summary of the dataset
- Datasets should be possible to open with a standard configuration of a common library (such as Pandas[12])

The most common identified indicators in the description (the ReadMe files) of the 20 top ranked repositories included:

- Links to basic concepts
- Links to resources
- Developer instructions / best practices
- Installation and processing instructions
- Mailing list / contact person / community
- Description of purpose

This prompts the question of whether dataset documentation such as ReadMe's should be facilitated in a more structured form, as has been contemplated in literature before. The facilitation of metadata provision through an interface (e.g. provided by CKAN) that prompts and facilitates the provision of the mentioned reuse indicators, while at the same time focusing on usability and user experience could alleviate some of the most obvious reuse barriers.

We further saw that high community engagement results in higher reuse which should incentivise portals to implement functionalities that allow such engagement. On the example of GitHub this includes the possibility to "follow" and "watch" a dataset as well as to "star" it. But this equally includes the ability to raise issues or allow targeted discussions around individual dataset. Providing an environment in which such community discussion can be attached to the dataset facilitates reuse. Feedback enables correction of data and can so increase data quality and value. This can be supported in a structured (e.g. feedback forms or pull requests) or unstructured way (forums, third party communication). Public feedback can also save time by making others aware that a dataset is unsuitable for a specific task.

We further suggest dedicated tracking of both user engagement in the form of, but not limited to:

- Dataset downloads

[12] https://pandas.pydata.org

- Dataset followers
- Dataset citations to track reuse more broadly

The more direct and accurate indication of actual dataset reuse can be acquired, the more accuracy the prediction model can gain as it is limited to the engagement proxies from which we derive reuse probabilities.

4. Conclusion

This work demonstrates the tension between calls for data reuse principles and actionable metrics and automated approaches facilitating data publishers and tools designers to implement functionalities supporting dataset reuse in an open collaborative environment. The findings point to a number of under-explored opportunities to encourage and facilitate dataset reuse on the web.

We show a potential direction to further develop both, guidance for dataset reuse, functionalities to predict a datasets reusability and at the same time recommend missing indicators to be added at the time of data publishing to enhance the value of existing datasets and enable meaningful reuse by wider audiences. It also allows platform developers or portal owners to focus tracking and capturing the right information from publishers to support better reusability of datasets.

This work could be built on by integrating functionalities that measure engagement with datasets in an automated way. Portals could support the automatic assessment of a dataset at the time of publication and recommend features that would increase reuse probability according to the proposed model. This would allow to increase a datasets reusability before publication, focusing on not just the data itself but also on documentation and other potentially relevant features of a project.

Even with current technologies, this approach can be used to inform system designers building functionalities to capture this information automatically; publishers in supplying certain information as metadata, and user experience designers, to inform the design of the interaction process between datasets reusers and the interface of a data portal

Portal owners can use this to inform their portal development, and open data users in the wider ecosystem can use these insights to help them identify the data sets that may be most useful to work with.

References

Akmon, D., Zimmerman, A., Daniels, M. and Hedstrom, M., 2011. The application of archival concepts to a data-intensive environment: working with scientists to understand data management and preservati

Koesten, L., Kacprzak, E., Tennison, J. and Simperl, E., 2019, May. Collaborative Practices with Structured Data: Do Tools Support What Users Need?. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-14).

Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H. and Hall, W., 2012. Linked open government data: Lessons from data. gov. uk. IEEE Intelligent Systems, 27(3), pp.16-24.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Daumeé III, H. and Crawford, K., 2018. Datasheets for datasets. arXiv preprint arXiv:1803.09010.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E. and Bouwman, J., 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data, 3.