Extra Material — Article "An Approach for Assessing Metadata Completeness in Open Data Portals

Juan Ribeiro Reis, Flavia Bernardini, José Viterbo
Higor dos Santos Pinto Carla Fernandes Torres

Aug 2020

1 Introduction

This technical report presents additional material of the paper An Approach for Assessing Metadata Completeness in Open Data Portals

Table 1 shows all the fields of the metadata schemas of the frameworks CKAN, Socrata and Opendatasoft.

Table 1: Main Frameworks Metadata Fields

Meta-data Schema	Label	Description
Ckan	Title	Allows intuitive labelling of the dataset for search, sharing and linking.
Ckan	Description	Additional information describing or analysing the data. This can either be static or an editable wiki which anyone can contribute to instantly or via admin moderation.
Ckan	Tags	See what labels the dataset in question belongs to. Tags also allow for browsing between similarly tagged datasets in addition to enabling better discoverability through tag search and faceting by tags.
Ckan	Unique identifier	Dataset has a unique URL which is customizable by the publisher.
Ckan	License	Instant view of whether the data is available under an open licence or not. This makes it clear to users whether they have the rights to use, change and redistribute the data.

Ckan	API key	Allows access every metadata field of the dataset and ability to change the data if you have the relevant permissions via API.
Ckan	Multiple formats (if provided)	See the different formats the data has been made available in quickly in a table, with any further information relating to specific files provided inline.
Ckan	Groups	Display of which groups the dataset belongs to if applicable. Groups (such as science data) allow easier data linking, finding and sharing amongst interested publishers and users.
Ckan	Data preview	Preview .csv data quickly and easily in browser to see if this is the dataset you want.
Ckan	Revision history	CKAN allows you to display a revision history for datasets which are freely editable by users (as is the-datahub.org).
Ckan	Extra fields	These hold any additional information, such as location data (see geospatial feature) or types relevant to the publisher or dataset. How and where extra fields display is customizable.
Opendatasoft	Title	The title of the dataset.
Opendatasoft	Description	The full description of the dataset (HTML is accepted).
Opendatasoft	Keywords	One or more keywords for the dataset, mostly used to make it easier to find in the portal.
Opendatasoft	Last modification	The last modification date of the dataset (manually set).
Opendatasoft	Publisher	The publisher of the dataset (the name of a person or of an organization).
Opendatasoft	Identifier	Technical identifier of the dataset.
Opendatasoft	License	The license attached to the dataset; should always be filled for any public dataset.
Opendatasoft	Geographic area	The geographical coverage of the data.
Opendatasoft	Language	The language (as a two-letter language code) of the datasets data and metadata.
Opendatasoft	References	One or more links to indicate the references or sources of the dataset.
Opendatasoft	Theme	One or more themes associated to the dataset (Environment...).
Opendatasoft	Attributions	Link of a source of the dataset that should be mentioned for legal reasons (e.g. if the license demands the mention of a specific source or organization).

Open-datasoft	Timezone	Forces the dataset visualizations to use the defined timezone for the date and datetime fields. It avoids the dataset visualizations to depend on the timezone on which the users computer is set.
Socrata	Title	Title helps users discover, select, and differentiate between similar datasets.
Socrata	Description	Description helps users discover, select, and differentiate between similar datasets.
Socrata	Tags	Tags link technical language, secondary categories, and acronyms to your dataset, aiding in user-executed searches.
Socrata	Last Updated	Last updated indicates of the recency of the data. Helps users determine usage of data.
Socrata	Contact Email	Consider including publicly-visible Contact Email on each dataset, which can be used by users to ask questions.
Socrata	Unique Identifer	A Unique Identifier is required for dataset management.
Socrata	Public Access Level	While most data on the platform will be public, Public Access Level gives us a means to track protected or sensitive data and provide a means for internal users to discover and access non-public data.
Socrata	Agency / Department	Responsible Agency/Department is helpful for navigation and to ensure a single responsible party.
Socrata	License / Rights	A License reduces legal uncertainty for data consumers or users.
Socrata	Geographic Unit	Geographic Unit indicates the geographic level at which the dataset is collected; also helps track the need to aggregate or summarize data.
Socrata	Temporal Coverage	Temporal Coverage provides an easy way to determine the value of a dataset.
Socrata	Data Dictionary	A Data Dictionary is essential to understanding how the data can be used. It can describe fields, differences between fields, and assess whether or not the data is appropriate for the intended use. Data Dictionaries could be published in both .csv and .pdf format.
Socrata	Permalink / Identifier	A Permalink helps provide continuity for accessing the dataset.
Socrata	Related Documents	Linking a Related Document provides the opportunity to include forms or other types of documents to help users understand the data. Not all datasets will have this information.

Socrata	Category	Category groups similar datasets together regardless of source and can be used to locate similar datasets.
Socrata	API Endpoint	An API Endpoint facilitates programmatic access to the data.
Socrata	Frequency of Data Change	Frequency - Data Change works together with the publishing frequency and helps set expectations for future updates as well as aids in planning.
Socrata	Frequency of Publishing	Frequency - Publishing works together with the Data Change frequency and helps set expectations for future updates as well as aids in planning.
Socrata	Public Access Level Comment	If the data is not public, consider providing an explanation and a means for people to access it if eligible.
Socrata	Data Steward	Consider including a Data Steward for each dataset to support the data coordinators and to answer dataset questions. This helps to track and triage data requests.
Socrata	Row Count	Row Count is a useful indicator of dataset size.
Socrata	Download URL	A Download URL provides access to the data for the purpose of open data.
Socrata	Link	A Link can provide more information on the origin of the dataset. Not all datasets will have this information.

Table 2 shows the metadata fields correspondences and their respective weights.

Table 2: Alpha Metadata Fields Correlation with European Portal Metadata Fields

Alpha Metadata Title	Europe Dataset Metadata Title	Weight
Description	Key: description	6.52%
Tags	Key: keywords	6.52%
Title	Key: title	6.52%
Unique identifier	Key: identifier	6.52%
API key	Key: owner_org	4.35%
Theme/Category/Groups	Key: concepts_eurovoc	3.26%
Theme/Category/Groups	Key: groups	3.26%
Last Updated	Key: modified_date	4.35%
Source	Key: resources(link)	4.35%
Related Documents	Key: resources	4.35%
Spatial Geographical Area	Key: geographical_coverage	4.35%
Contact Email	Key: contact_email	2.17%
Extra fields	Key: extras	2.17%
Agency/Department	Key: organization	2.17%

Data Dictionary	Key: resources(name)	2.17%
Data preview	-	2.17%
Download URL	Key: resources(download)	2.17%
Frequency	Key: temporal_granularity	1.09%
Frequency	Key: accrual_periodicity	1.09%
Permalink/Identifier	Key: url	2.17%
Language	Key: language	2.17%
Format	Key: resources(format)	2.17%
Public Access Level	Key: private	2.17%
Public Access Level Comment	-	2.17%
Revision history	Key: revision_timestamp	2.17%
Row Count	-	2.17%
Timezone	-	2.17%
License	Key: license_id	2.17%
License	Key: license_title	2.17%
License	Key: license_url	2.17%
Data Steward	Key: maintainer	2.17%
Publisher	Key: author	2.17%
Temporal Coverage	Key: temporal_coverage_to	1.09%
Temporal Coverage	Key: temporal_coverage_from	1.09%
Notcorrelated	Key: contact_name	0.00%
Notcorrelated	Key: contact_webpage	0.00%
Notcorrelated	Key: contact_address	0.00%
Notcorrelated	Key: alternative_title	0.00%
Notcorrelated	Key: capacity	0.00%
Notcorrelated	Key: contact_telephone	0.00%
Notcorrelated	Key: creator_user_id	0.00%
Notcorrelated	Key: interoperability_level	0.00%
Notcorrelated	Key: isopen	0.00%
Notcorrelated	Key: metadata_created	0.00%
Notcorrelated	Key: metadata_language	0.00%
Notcorrelated	Key: metadata_modified	0.00%
Notcorrelated	Key: name	0.00%
Notcorrelated	Key: num_resources	0.00%
Notcorrelated	Key: num_tags	0.00%
Notcorrelated	Key: owner_org	0.00%
Notcorrelated	Key: rdf	0.00%
Notcorrelated	Key: relationships_as_object	0.00%
Notcorrelated	Key: relationships_as_subject	0.00%
Notcorrelated	Key: release_date	0.00%
Notcorrelated	Key: revision_id	0.00%
Notcorrelated	Key: state	0.00%
Notcorrelated	Key: status	0.00%

Notcorrelated	Key: tracking_summary	0.00%
Notcorrelated	Key: type	0.00%
Notcorrelated	Key: type_of_dataset	0.00%
Notcorrelated	Key: version	0.00%
Notcorrelated	Key: version_description	0.00%
Notcorrelated	Key: maintainer_email	0.00%
Notcorrelated	Key: author_email	0.00%

Table 3 shows the metadata fields correspondences and their respective weights.

Table 3: Alpha Metadata Fields Correlation with NYS Metadata Fields

Metadata Title	Dataset Metadata	Weight
Description	Description	6.52%
License	-	6.52%
Tags	Keywords	6.52%
Theme/Category/Groups	Category	6.52%
Title	Name	6.52%
Unique identifier	U ID	6.52%
API key	api_endpoint	4.35%
Last Updated	Last Update Date (data)	4.35%
Related Documents	-	4.35%
Source	Source Link	4.35%
Agency/Department	Agency	2.17%
Contact Email	Contact Information	2.17%
Data Dictionary	-	2.17%
Data preview	Derived View	2.17%
Data Steward	-	2.17%
Download URL	-	2.17%
Extra fields	-	2.17%
Format	-	2.17%
Frequency	Posting Frequency	2.17%
Language	-	2.17%
Permalink/Identifier	URL	2.17%
Public Access Level	-	2.17%
Public Access Level Comment	-	2.17%
Publisher	Data Provided By	2.17%
Revision history	-	2.17%
Row Count	-	2.17%
Spatial Geographical Area	Coverage	2.17%
Spatial Geographical Area	Localities	2.17%
Temporal Coverage	-	2.17%

Timezone	-	2.17%
Notcorrelated	Type	0.00%
Notcorrelated	Domain	0.00%
Notcorrelated	Organization	0.00%
Notcorrelated	See Also	0.00%
Notcorrelated	Granularity	0.00%
Notcorrelated	Limitations	0.00%
Notcorrelated	Notes	0.00%
Notcorrelated	Owner	0.00%
Notcorrelated	Visits	0.00%
Notcorrelated	Downloads	0.00%
Notcorrelated	Creation Date	0.00%
Notcorrelated	Parent UID	0.00%
Notcorrelated	County Filter	0.00%
Notcorrelated	County Column	0.00%
Notcorrelated	Municipality Filter	0.00%
Notcorrelated	Municipality_Column	0.00%

Table 3 shows the Beta Metadata Fields Correlation With European Portal Metadata Fields.

Table 4: Beta Metadata Fields Correlation With European Portal Metadata Fields

Beta Metadata Title	Europe Dataset Metadata Title	Weight
Description	Key: description	5.13%
Tags	Key: keywords	5.13%
Title	Key: title	5.13%
Unique identifier	Key: identifier	5.13%
Last Updated	Key: modified_date	3.85%
Related Documents	Key: resources	3.85%
Source	Key: resources(link)	3.85%
Spatial Geographical Area	Key: geographical_coverage	3.85%
Theme/Category/Groups	Key: concepts_eurovoc	2.56%
Theme/Category/Groups	Key: groups	2.56%
API key	Key: owner_org	2.56%
Contact Email	Key: contact_email	2.56%
Download URL	Key: resources(download)	2.56%
Format	Key: resources(format)	2.56%
Language	Key: language	2.56%
Permalink/Identifier	Key: url	2.56%
Publisher	Key: author	2.56%
License	Key: license_id	1.71%
License	Key: license_title	1.71%

License	Key: license_url	1.71%
Access URL	-	1.28%
Agency/Department	Key: organization	1.28%
Byte size	-	1.28%
Checksum	-	1.28%
Conforms to	-	1.28%
Data Dictionary	Key: resources(name)	1.28%
Data preview	-	1.28%
Data Steward	Key: maintainer	1.28%
Documentation	-	1.28%
End date/time	-	1.28%
Extra fields	Key: extras	1.28%
Frequency	Key: temporal_granularity	1.28%
Frequency	Key: accrual_periodicity	1.28%
Linked schemas	-	1.28%
Media type	-	1.28%
Provenance	Key: owner_org	1.28%
Public Access Level	Key: private	1.28%
Public Access Level Comment	-	1.28%
Release date	Key: release_date	1.28%
Revision history	Key: revision_timestamp	1.28%
Row Count	-	1.28%
Source	-	1.28%
Start date/time	-	1.28%
Temporal Coverage	Key: temporal_coverage_to	1.28%
Temporal Coverage	Key: temporal_coverage_from	1.28%
Timezone	-	1.28%
Type	Key: type	0.64%
Type	Key: type_of_dataset	0.64%
Version	Key: version	0.64%
Version	Key: version_description	0.64%
Notcorrelated	Key: contact_name	0.00%
Notcorrelated	Key: contact_webpage	0.00%
Notcorrelated	Key: contact_address	0.00%
Notcorrelated	Key: alternative_title	0.00%
Notcorrelated	Key: capacity	0.00%
Notcorrelated	Key: contact_telephone	0.00%
Notcorrelated	Key: creator_user_id	0.00%
Notcorrelated	Key: interoperability_level	0.00%
Notcorrelated	Key: isopen	0.00%
Notcorrelated	Key: metadata_created	0.00%
Notcorrelated	Key: metadata_language	0.00%
Notcorrelated	Key: metadata_modified	0.00%

Notcorrelated	Key: name	0.00%
Notcorrelated	Key: num_resources	0.00%
Notcorrelated	Key: num_tags	0.00%
Notcorrelated	Key: rdf	0.00%
Notcorrelated	Key: relationships_as_object	0.00%
Notcorrelated	Key: relationships_as_subject	0.00%
Notcorrelated	Key: revision_id	0.00%
Notcorrelated	Key: state	0.00%
Notcorrelated	Key: status	0.00%
Notcorrelated	Key: tracking_summary	0.00%
Notcorrelated	Key: maintainer_email	0.00%
Notcorrelated	Key: author_email	0.00%

The metadata and all correspondences and their respective weights can be seen in Table 5.

Table 5: Beta Metadata Fields Correlation with NYS Metadata Fields

Metadata Title	Dataset Metadata	Weight
Description	Description	5.13%
License	-	5.13%
Tags	Keywords	5.13%
Theme/Category/Groups	Category	5.13%
Title	Name	5.13%
Unique identifier	U ID	5.13%
Last Updated	Last Update Date (data)	3.85%
Related Documents	-	3.85%
Source	Source Link	3.85%
Spatial Geographical Area	Coverage	1.92%
Spatial Geographical Area	Localities	1.92%
API key	api_endpoint	2.56%
Contact Email	Contact Information	2.56%
Download URL	-	2.56%
Format	-	2.56%
Frequency	Posting Frequency	2.56%
Language	-	2.56%
Permalink/Identifier	URL	2.56%
Publisher	Data Provided By	2.56%
Temporal Coverage	-	2.56%
Access URL	-	1.28%
Agency/Department	Agency	1.28%
Byte size	-	1.28%
Checksum	-	1.28%

Conforms to	-	1.28%
Data Dictionary	-	1.28%
Data preview	Derived View	1.28%
Data Steward	-	1.28%
Dataset distribution	-	1.28%
Documentation	-	1.28%
End date/time	-	1.28%
Extra fields	-	1.28%
Linked schemas	-	1.28%
Media type	-	1.28%
Provenance	-	1.28%
Public Access Level	-	1.28%
Public Access Level Comment	-	1.28%
Release date	Creation Date	1.28%
Revision history	-	1.28%
Row Count	-	1.28%
Start date/time	-	1.28%
Timezone	-	1.28%
Type	Type	1.28%
Version	-	1.28%
Notcorrelated	Domain	0.00%
Notcorrelated	Organization	0.00%
Notcorrelated	See Also	0.00%
Notcorrelated	Granularity	0.00%
Notcorrelated	Limitations	0.00%
Notcorrelated	Notes	0.00%
Notcorrelated	Owner	0.00%
Notcorrelated	Visits	0.00%
Notcorrelated	Downloads	0.00%
Notcorrelated	Parent UID	0.00%
Notcorrelated	County Filter	0.00%
Notcorrelated	County Column	0.00%
Notcorrelated	Municipality Filter	0.00%
Notcorrelated	Municipality_Column	0.00%