FAIR self assessment for project: National Australian Twitter Collection

Completed 7 Oct. 2019

	Questions for each FAIR component ↓	Answer options: Increasingly FAIR -->				
	FINDABLE					
Q1	Does the dataset have any identifiers assigned?	No identifier	Local identifier	Web address (URL)	Globally unique, citable and persistent identifier (e.g. DOI, PURL, or Handle)	
A1	Start of project				All individual tweets and account profiles are identified by the unique numerical IDs assigned to them by Twitter. Users of the NATC dataset can resolve these to URLs showing the current, live state of these entities (which may differ from the state at the point of gathering if Twitter users subsequently delete their tweets, or if Twitter accounts change their names or profile information, or are suspended or deleted). It is common practice in the scholarly literature to cite such content much like any other Web content: with reference to this URL and the time at which they were last visited by the researcher.	
	End of project				As above.	
	Two years' time				As above.	
Q2	Is the identifier included in all metadata records or metadata files describing the data?	No	Yes			
A2	Start of project		Yes, these identifiers are an inherent element of the dataset.			
	End of project		As above.			
	Two years' time		As above.			
Q3	Is the data described by a metadata record?	The data is not described	Brief title and description	Brief title and description, and multiple other fields filled out, albeit briefly.	Comprehensively (a min metadata template will be provided) using a formal machine-readable metadata schema.	
A3	Start of project			Current metadata records and related information are inherited from the TrISMA ARC LIEF project and further enhanced by the QUT Digital Observatory. A more comprehensive metadata record, including information on coverage limits and usage encumbrances, has yet to be developed, and will need to evolve in line with the establishment of operational and accreditation guidelines as the NATC transitions to a continuing national entity.		
	End of project			As above.		
	Two years' time				Comprehensive metadata record reflecting operational and access frameworks then established by the NATC entity.	
Q4	What type of repository or registry is the metadata record in?	The data is not described in any registry or repository	Local institutional repository	Domain-specific repository	Generalist public repository	Data is in one place but discoverable through several places (i.e. other registries, RDA, Google Data Search)
A4	Start of project		Existing version of the dataset included in QUT's institutional repository.			
	End of project		As above.			

	Two years' time					The overall NATC dataset will be discoverable through multiple registries. Direct indexing of the data by search engines will never be possible, however, as this would violate Twitter's terms of service as well as users' reasonable expectations of data privacy.
	ACCESSIBLE					
Q5	How accessible is the data? Note: The access method (s) must be explicitly stated in the metadata record, e. g. if any authentication is needed, or there are any restrictions to access.	No metadata record	Access to metadata only	Unspecified access conditions e.g. "contact the data custodian to discuss access"	Embargoed access after a specified date; or A deidentified version of the data is publicly accessible	Fully accessible public, or to persons who meet and follow explicitly stated conditions and processes, e.g. ethics approval for sensitive data
A5	Start of project					Fully accessible to researchers at participating institutions who have received appropriate ethics clearances from their home institutions, subject to the limitations stipulated to such clearances.
	End of project					As above.
	Two years' time					Fully accessible to researchers at participating institutions who have received appropriate ethics clearances from their home institutions, subject to the limitations stipulated to such clearances.
Q6	Is the data available online without requiring specialised protocols or tools once access has been approved?	No access to data	By individual arrangement	File download from online location	Non-standard web service (e.g. OpenAPI/Swagger/informal API)	Standard web service API (e.g. OGC)
A6	Start of project				Data hosted on dedicated database; access passwords provided and managed by QUT Digital Observatory. Data access through direct database queries in SQL flavours, or in the form of CSV and equivalent exports.	
	End of project				As above.	
	Two years' time					Data hosted on AWS Athena database; access managed by NATC entity. Data access through direct database queries in Athena SQL flavours, and via research tools and languages (Tableau, Python, R, etc.) that support Athena.
Q7	Does the repository/registry agree to maintain the persistence of the metadata record, even if the data product is no longer available?	No (or not applicable, if no metadata record exists)	Unsure	Yes		
A7	Start of project	Not applicable: NATC is designed as a continuously updated, live dataset.				
	End of project	As above.				
	Two years' time	As above, with additional repository at the National Library				
	INTEROPERABLE					
Q8	Are the data available in (an) open (file) format(s)?	Data are mostly available only in a proprietary format	Data are available in an open format	Data are available in an open, documented, widely-used standard format (i.e. NetCDF, CSV, JSON, XML, etc)		
A8	Start of project			Data available for querying in dedicated database. Database exports available in CSV and equivalent formats.		
	End of project			As above.		

	Two years' time			Data hosted on AWS Athena database; access managed by NATC entity. Data access through direct database queries in Athena SQL flavours, and via research tools and languages (Tableau, Python, R, etc.) that support Athena. Data export in CSV and other formats deprecated except for limited subsets and special use cases in order to ensure data security and privacy.		
Q9	Are the data machine readable?	The data are unstructured	The data are structured and machine-readable (i.e. csv, JSON, XML, RDF, database files, etc)			
A9	Start of project		Data stored in dedicated relational database and machine-readable for accredited users.			
	End of project		As above.			
	Two years' time		As above.			
Q10	What best describes the types of vocabularies/ontologies/tagging schemas used to define the data elements?	Data elements are not described (i.e. fields or objects are labelled with codes or not at all)	Data elements are described (so that a human user can correctly interpret the data), but no standards have been used in the description	Recognised standards have been used in the description of data elements, but no published vocabularies with resolvable URIs are used	Published vocabularies using resolvable identifiers linking to explanations are used, so that the data can be read and understood by machines as well as humans.	Published vocabularies using **persistent** resolvable identifiers linking to explanations are used, so that the data can be read and understood by machines as well as humans.
A10	Start of project			Data ontologies inherited from Twitter data structures, as the source of the data. Twitter's own vocabularies are readily available as part of overall Twitter developer information.		
	End of project			As above.		
	Two years' time				Further enhancements to dataset documentation provide more explicit references and links to Twitter data vocabularies, complemented by training materials that introduce these concepts to non-developers.	
Q11	How is the relationship to other data and resources (e.g. related datasets, services, publications, etc) described in the metadata, to provide context around the data?	There are no links to other metadata or data	The metadata record includes URI links to related metadata, data and definitions	Qualified links to other resources are recorded in a machine readable format, e.g. a linked data format such as RDF		
A11	Start of project	The specific nature of this dataset means that this question does not apply particularly clearly. All tweet and account information contained in the dataset uses the unique numerical IDs assigned by Twitter, which can be readily converted into URLs showing the current, live state of these entities. These may differ from the state at the point of gathering if Twitter users subsequently delete their tweets, or if Twitter accounts change their names or profile information, or are suspended or deleted. Such IDs may also be used to establish relationships between the NATC dataset and other resources that represent Twitter data. Beyond this, however, the potential uses of the NATC datasets are so diverse that users are likely to establish unique and project-specific relationships with other datasets that would be impossible to anticipate in the provision of further metadata.					
	End of project	As above.				

	Two years' time	As above.				
	REUSABLE					
Q12	Which of the following best describes the license (usage rights) attached to the data?	No license is applied	Non-standard license applied, without a license deed URL encoded in a machine-readable format (e.g. RDF/XML) in the metadata record	Non-standard license applied, WITH the license deed URL encoded in a machine-readable format (e.g. RDF/XML) in the metadata record	Standard license applied (e.g. Creative Commons), without a license deed URL encoded in a machine-readable format (e.g. RDF/XML) in the metadata record	Standard license applied (e.g. Creative Commons), WITH the license deed URL encoded in a machine-readable format (e.g. RDF/XML) in the metadata record
A12	Start of project		In light of the provenance of the data from a commercial social media platform, and the reuse and privacy considerations that apply to these data, existing datasets have implemented a dedicated user accreditation process that also stipulates usage restrictions. This information is not currently encoded in machine-readable format.			
	End of project		The legal advice received for this project has identified further opportunities for improving and clarifying the accreditation and access process, including suggestions for revisions to the licence information provided to researchers. These will be implemented once the legal advice has been finalised.			
	Two years' time		Revised accreditation and access frameworks implemented by NATC entity, and licence terms encoded in machine-readable format.			
Q13	How much provenance information has been captured to facilitate data reuse? i.e. project objectives, data generation/collection (including from external sources) and processing workflows.	No provenance information is recorded	Partially recorded	Comprehensively recorded in a text format (i.e. TXT or PDF)	Comprehensively recorded in a machine readable format (i.e. in metadata record's schema or PROV, or in RDF, JSON, NetCDF, XML, etc)	
A13	Start of project		Overall information on dataset provenance, format, and limitations provided to researchers only in high-level detail. Background on data collection covered in a number of published, peer-reviewed, citable publications.			
	End of project		As above.			
	Two years' time				Full details on dataset provenance, format, and limitation provided in extensive detail, available in human- as well as machine-readable formats.	