Research Data Management Policy & Implementation Strategy

Tomorrow's CITIES Urban Risk in Transition
UK Research and Innovation GCRF
Tomorrow's Cities is the UKRI GCRF Urban Disaster Risk Hub

Tomorrow's Cities Research Data Management Policy & Implementation Strategy

Purpose: ***To enable a comprehensive and co-created approach to Research Data Management with UK/DAC partners and researchers which will align and enhance Tomorrow's Cities (TC) mission and impact.***

Data management goals

Tomorrow's Cities has three principal goals for research data management:

- to support each city in developing its own open data infrastructure, following the roadmap in "Open Data Infrastructure for City Resilience" (UNISDR 2018);
- to create a data hub outside the Hub, both to support continuity and disaster recovery and to provide a common research environment for the Hub;
- to facilitate data sharing and research pooling across the consortium, but particularly "south-south partnerships".

Data use in Tomorrow's Cities

The Hub will create, collect and collate data of multiple types from multiple sources centred in the four global cities, for a number of research purposes:

- detailed local mapping of hazards and vulnerability, incorporating local knowledge;
- real-time monitoring of hazards, their impacts and people's responses;
- initialisation of predictions and models (both manual and computer-based);
- verification and improvement of predictions and model (to build trust in their accuracy);
- characterisation of multi-hazard events;
- assessing and improving risk management systems (hazard, impact, response);
- reassessment and reuse of the research.

TC will follow the FAIR data principles of findability, accessibility, interoperability and reusability (cf. doi:10.1038/sdata.2016.18):

- **findable**: all project data will be described in a common, Web-accessible catalogue;
- **accessible**: all project data will be open by default, restricted only if necessary. Data will be made accessible on the Web by direct download, API and portal;
- **interoperable**: data will be both stored and presented in open-standard formats;
- **reusable**: by default data will be published under a rights waiver (public domain), or, where licences are needed, under the most open possible from Creative Commons 4.0.

Data collection and hosting

To create a common research environment, the policy goals will promote inter-city research and provide off-site continuity for city data stores, and collected data is hosted in a "data hub" at the University of Edinburgh's Advanced Computing Facility. The data hub will be built on CKAN (https://ckan.org /) and will be used to curate data actively over the TC lifetime. Storage will come from the World-Class Data Infrastructure. Historical "background" data will either be ingested, if sensible, or referenced remotely, making the data hub a "virtual data lake". Our policy is to engender a major effort to put into a common catalogue of descriptive metadata for both "local" and "remote" data resources.

Data areas and data types

The dynamic and flexible nature of TC means our data is represented by a wide array of types, volumes and rates. The types of data we expect to be managed through implementation of this policy will include:

- remotely sensed data from ground and space, giving consistency over long periods, common spatial and temporal templates, and conveniently packaged data delivery;
- *in situ* instruments, whether on stationary or mobile platforms (comparability between instruments is crucial so equipment and observing standards must be observed and metadata delivered, especially difficult for mobile data sources);

- i*n situ* reports of hazards, their impact, and of people's responses, whether from instruments or human observation (standards specification and reporting of metadata remain important even for qualitative value);
- surveys of the receipt and reaction to warnings, notably the choice of media and language (more in depth social science tools, such as focus groups and interviews, etc);
- model analyses and predictions, both the outputs that might be used to inform advice and warnings, and the conditions needed to rerun the event for research;
- communications that form part of the emergency management process, including forecasts, their interpretation, advice, warnings, supporting evidence, blogs, social media posts, media presentations;
- research information, such as survey templates, questionnaires, model specifications, standards, format specifications, sampling maps etc.

Standards and metadata

Our approach is to follow best practice in using open standards for recording data. As a reference point we will use https://earthdata.nasa.gov/user-resources/standards-and-references . The CKAN open-source data repository platform is in wide use in open data initiatives around the world, including sites in Africa, Asia and Latin America (cf. https://ckan.org/about/instances/). CKAN has well-defined methods for federation across instances, making it an excellent choice for wider interoperability and data sharing.

Relationship to other data available in public repositories

Following the “virtual data lake” model the Hub will connect to existing data resources within the four cities and beyond, as and when required by the research colleges. Some data resources will be provided by city partners under restricted access conditions and will be managed using an appropriate authorisation regime.

Secondary use

We intend that data resources built up during the Hub’s project life (“foreground data”) will be freely available both during and beyond the project lifetime. Where geoscience data are generated we will deposit them with the UK National Geoscience Data Centre, following NGDC’s Ingestion Policy and “good data deposit guidelines.”

Methods for data sharing

The Hub data will be available through the data hub under a rights waiver (public domain) where possible.

Proprietary data

Data will be open by default. Data will be restricted if necessary only for reasons of personal or cultural sensitivity, and in these case mechanisms will be sought to release aggregated or de-identified versions.

Timeframes

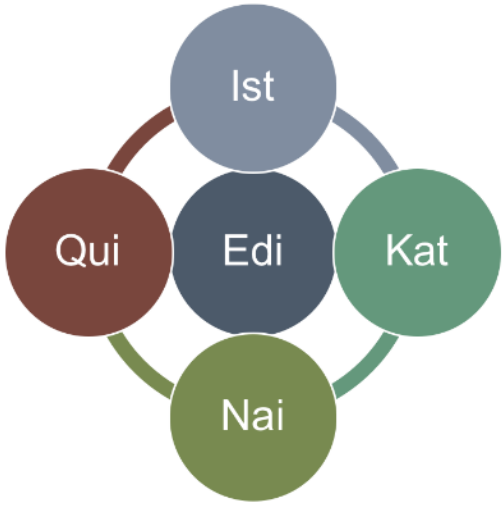
Data will be freely available on the data hub as soon as they have been curated.

Format of the final dataset

By following the FAIR principles and ensuring data are curated in open formats throughout the project our policy will aim to ensure that final products are widely reusable.

Implementation Strategy

1. Set up central Research Data Store in Edinburgh.
2. Develop tailored data management plans co created with each DAC city.
3. Set up a Research Data Store in each DAC city.
4. Connect city RDSs to the central RDS.
5. Synchronise data between city RDS and central RDS following agreed policies.



Research Data Store?

Our starting recipe for an RDS is:

- a modest computer,
- with some data storage space,
- hosting a CKAN data catalogue,
- with network connections to the Edinburgh Hub.

Modest computer?

The RDS computer should be powerful enough to run the CKAN software (CKAN suggests 4 CPU cores, 8 GB memory, 500 GB disk storage).

Data storage space?

Additional storage (beyond the 500 GB for CKAN itself) will be needed for the local data store.

CKAN data catalogue?

The implementation strategy follows the CKAN[1] i "The world's leading Open Source data portal platform" (their website quote) approach and is used in a lot of open data and open government data portals worldwide. CKAN will install on version 16.04 or later of the free, open source Ubuntu Linux operating system. University of Edinburgh have developed tools to make installation and configuration as easy as possible on other Linux flavours.

Network connections to University of Edinburgh – Implementation Considerations ?

The University of Edinburgh will connect each of the DAC city RDS catalogues to the Edinburgh one using a CKAN feature called "metadata harvesting". Thereafter the University of Edinburgh will also offer the central RDS as an archive, backup or second "live" site to all the DAC city RDSs – to be determined through individual DAC city DMPs.

Ingesting data and metadata

Simple ingest through CKAN

CKAN is very flexible and extending for implementation and it is new functionality is relatively easy. By default the set it will be implemented with data upload and metadata description functions, meaning small datasets (up to a few gigabytes) can be uploaded directly into CKAN and stored behind the scenes.

For every dataset uploaded a basic metadata record will be made, a description of the dataset, who uploaded it etc. The Hub will adopt the DCAT metadata profile (with a couple of extensions) as the *minimum* metadata record to enable TC researchers to find and re-use each dataset. Because subject areas vary, this minimum record will not have much detailed

[1] https://ckan.org/

scientific information in it. CKAN is highly extensible in the metadata it supports, though, and these details will be captured in the individual city data management plans.

Ingesting larger datasets

Ingesting large datasets (multi-gigabyte files, for instance) directly into CKAN doesn't make sense for implementation. Instead, a metadata record can be created that just points to a file somewhere in the RDS storage system. Synchronising large datasets between a DAC city RDS and Edinburgh will be planned as needed for each city.

Capacity planning

University of Edinburgh will build the central RDS on top of a wider chunk of data infrastructure funded by the UK and Scottish Governments (the "World-Class Data Infrastructure"). This is a multi-petabyte facility and so central capacity for Tomorrow's Cities will not be a problem.

Individual cities may not have enough local capacity to store some of the larger kinds of dataset expected (broadband seismometer traces, drone images…). In these cases the Edinburgh team will work with the DAC city team to stream as much data as possible "live" straight to the Edinburgh RDS.

Accessing data

Open versus restricted data

As noted, one of the key goals of the Hub data stores is to create as much open data as possible (we've promised our funders this). However, this may not always be possible, or it may be desirable to wait a period of time before opening some data up to the outside world. The minimum metadata model for Hub data records includes a *DataTag* field which indicates the openness (or not) of individual datasets. DataTags are traditionally coloured, with blue indicating totally open, and levels of sensitivity rising through green, yellow and red.

Download from CKAN

For smaller datasets (around a few gigabytes) CKAN supports direct download of accessible data through the Web browser. For larger datasets, individual access arrangements will need to be made, perhaps using other tools like ftp.

Personal data policies

The Hub expects to collect a fair amount of personal, confidential data on a regular basis (notably surveys or some citizen science outputs). Our policy will adopt a different policy to handling personal data (this is reflected in a separate "sensitive data" data management plan template):

- Identifiable personal data ("confidential data") MUST NOT leave their country of origin. The Edinburgh hub SHALL NOT harvest or back up confidential data.
- De-identified (or "pseudonymised") personal data MAY leave their country of origin (if allowed by that country's data processing laws). They MUST be tagged 'green' or higher, and a Data Processing Agreement between the Data Controller and the receiving party (the University of Edinburgh in the case of the Edinburgh Hub) MUST be completed.
- Data dictionaries that relate response ids in pseudonymised datasets to individuals (i.e. the "pseudonymisation keys") MUST be tagged 'yellow' or higher. These must never be transported over the Internet, MUST be encrypted and MUST NOT be stored on an Internet-accessible computer.
- Aggregated personal data MAY leave their country of origin (if allowed by that country's data processing laws). They MUST be tagged 'green' or higher, and a Data Processing Agreement between the Data Controller and the receiving party (the University of Edinburgh in the case of the Edinburgh Hub) MUST be completed.

- Minimum metadata (as per the Appendix) can be defined as ‘blue’ data, and as such SHOULD always be made public.
- Collection level metadata, and metadata relating to aggregated datasets, SHOULD always be made public.
- Aggregated, or otherwise anonymised, data MAY be made available through publicly available endpoints. Data at a finer granularity (i.e. series or object level) SHOULD, under normal circumstances, not be publicly available. Requests for access to this data should be made to the Data Owner or Data Curator, as appropriate. It is likely that data released will have a restricted licence, and to have other restrictions placed on it to comply with local and international laws.
- All ‘blue’ metadata SHOULD be harvested by the Edinburgh-hosted CKAN instance.

Appendix 1: Mandatory Metadata Fields

The list below contains the minimum metadata requirements of every dataset in the Hub.

Note the Hub has not specified metadata for data objects, as it is recognised that metadata requirements at this granularity will be highly dependent upon the nature of the data. Metadata relating to the catalogue itself will be maintained by the catalogue publishers.

The Hub will use Version 2 of the DCAT ("*Data Catalog*") Vocabulary[2] as a base to describe datasets. This requires all datasets to be instances of dcat:Dataset.

DCAT Property	Definition	Notes
title	Title of the dataset	This should be as descriptive as possible. A longer description could be added as a dcat:description property
identifier	Unique identifier of the dataset	This can be any valid string. It could be a DOI for the dataset.
spatial	The geographical area covered by the dataset	This can either be exact (i.e. a Latitude / Longitude) or a place name
temporal	The temporal period that the dataset covers	The catalog maintainers will ensure that the catalog software formats this correctly
theme	A main category of the resource	Each dataset can have many themes. The catalog maintainers will ensure that the themes are compliant with the DCAT documentation
publisher	The entity responsible for making the item available	This could just be a name of a person or organization

Additional Properties	Definition	Notes
datatag	A tag representing the sensitivity and handling requirements of the data object.	Note that, if a data object has a DataTag, it has one and one only, regardless of how many possible classification bases might be applied (e.g. for an object subject to multiple regulatory frameworks). Which tag the object should carry may be a matter of policy, but the path of greatest risk reduction suggests that the strictest tag suggested should be the one used

The Hub has not specified any further optional properties, but acknowledges further properties should be used to describe datasets more richly. Properties used can be taken from any relevant vocabulary, although please ensure that both *domain* and *range* are valid, where applicable, and that the dataset is a valid instance of the Class that the vocabulary defines. Help and advice can be provided by the catalog maintainers.

[2] https://www.w3.org/TR/vocab-dcat-2/

Appendix 2: DataTags Colour Coding

The Hub will use a colour-coded Data Tag approach for risk-based classification of potentially sensitive data. This follows an approach developed by DANS in the Netherlands around GDPR (see Doorn & Thomas, *Tagging Privacy-Sensitive Data According to the New European Privacy Legislation: GDPR DataTags - a Prototype,* International Digital Curation Conference (IDCC), 2018), based on the original work at Harvard by Sweeney, Crosas & Bar Sinai (2015)[3]. The first prototype was developed under the EUDAT2020 project using the Zingtree decision tree application to support researchers in complying with the GDPR.

Risk Class	Technical and organisational measures	GDPR category	Description
0 – Public, blue tag	None.	Non-personal data.	Dataset contains no information that refers to any identified or identifiable living individual.
1 – Basic, green tag	Although anonymised data are out of scope of the GDPR, protection and authentication are desirable since de-identification is always possible. In addition, in aggregation with other datasets, the data could be traced back to original. Registration necessary, processing agreement is required for Edinburgh, resulting in demonstrable accountability.	Anonymised personal data.	The dataset does contain personal information, but the researcher has made sure that this data is anonymised. Principles of anonymisation have been followed accordingly.

II – Increased, yellow tag	Examples include, but are not limited to:
	• Processing agreement
	• Data minimisation
	• Pseudonymisation
	• Authentication access policy:
	  • registered users only
	  • mandatory identification
	  • depositor approval
		Personal data. Consent obtained, including child's consent.	Dataset contains personal data. This data is collected in a lawful manner on the basis of obtained consent. This consent is obtained in compliance with articles 5, 6, and 7, and 8 in case of data subjects below 16.
			Message: “continue, but make sure appropriate safeguards are in place.”

III – High, red tag	Examples include, but are not limited to: • Processing agreement • Data minimisation • Pseudonymisation • Encryption • Two- or multi-factor authentication • Authentication access policy (depositor approval):	Special categories of personal data (consent obtained, including child's consent). Unconsented personal data.	Given the answers provided, special categories of personal data are expected to be processed. This data is collected in a lawful manner on the basis of obtained consent. Since the GDPR provides multiple articles dedicated to these categories,

[3] http://datatags.org/

	• registered users only • protected environment access (special permission only) • mandatory identification • depositor approval		additional prudence is advised. Message: “continue, but make sure appropriate safeguards are in place.”