Linking datasets with user commentary, annotations and publications: the CHARMe project

Jon Blower
j.d.blower@reading.ac.uk
University of Reading
On behalf of all CHARMe partners!

http://www.charme.org.uk

CHARMe

(Jan 2013 – Dec 2014)

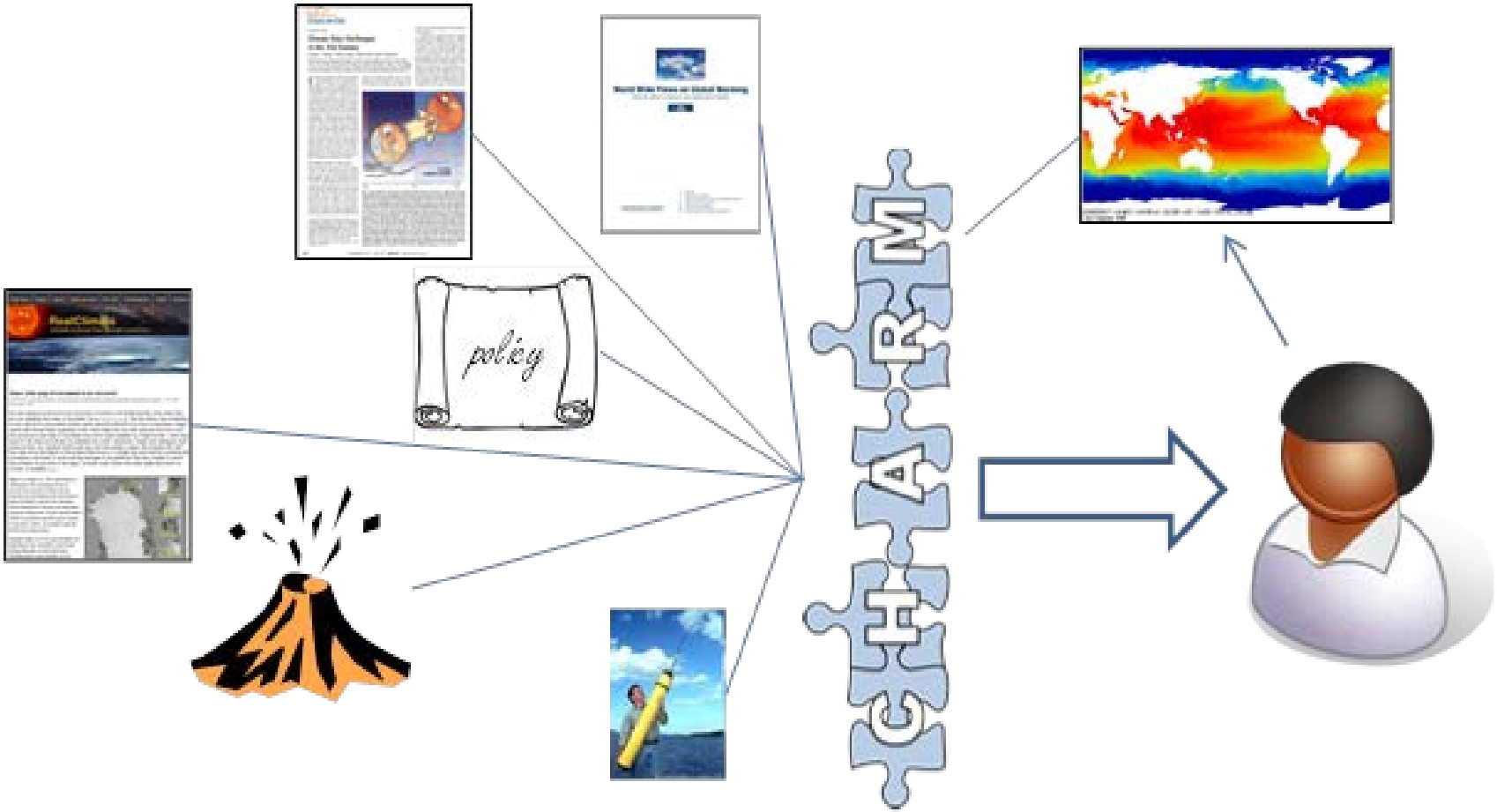
"CHARacterization of Metadata to enable high-quality climate applications and services"

How can climate data users decide whether a dataset is fit for their purpose?

(N.B. We consider that "data quality" and "fitness for purpose" are the same thing)

Not specific to climate data!

“Commentary metadata”



Examples of commentary metadata

- **Post-fact annotations**, e.g. citations, ad-hoc comments and notes;
- **Results of assessments**, e.g. validation campaigns, intercomparisons with models or other observations, reanalysis;
- **Provenance**, e.g. dependencies on other datasets, processing algorithms and chain, data source;
- **Properties of data distribution**, e.g. data policy and licensing, timeliness (is the data delivered in real time?), reliability;
- **External events** that may affect the data, e.g. volcanic eruptions, El-Nino index, satellite or instrument failure, operational changes to the orbit calculations.

General rule: information originates from **users or external entities**, not original data providers

  - However, sometimes information is not available from the data provider!

Primary use case

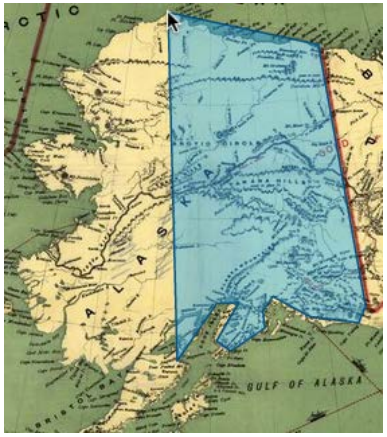
1. User searches data archive for relevant datasets
2. Each dataset in the results has two “CHARMe buttons” for reading and creating commentary metadata about the dataset
3. Pressing the button brings up pop-up listing all the annotations about that dataset.

Very much like METAFOR / ES-DOC system (right) for climate model descriptions

(Can be implemented with very little impact on **existing websites**, using Javascript magic)

Other use cases

- Viewing “significant events” in timeseries data (cf. Google Finance)
- Creating and discovering annotations about dataset subsets (cf. maphub.github.io)
- Enabling intercomparisons of data and metadata (cf. ES-DOC)



Open Annotation

- We propose to use Open Annotation (W3C standard) for modelling annotations
- Based on Linked Data technologies
  - RDF, SPARQL etc
  - Used by data.gov.uk, Australian Bureau of Meteorology, UK Met Office, many more!
- Data model is simple and flexible
  - We don't have to design a rigid schema or object model up-front
  - Can be added to as time goes on
- Can record the *motivation* behind an annotation
  - Bookmarking, classifying, commenting, describing, editing, highlighting, questioning, replying… (lots more)
  - Covers a lot of CHARMe use cases!
- An annotation can have multiple targets
  - Another CHARMe requirement
- There is even (limited) support for annotating subsets of resources
  - An advanced CHARMe requirement

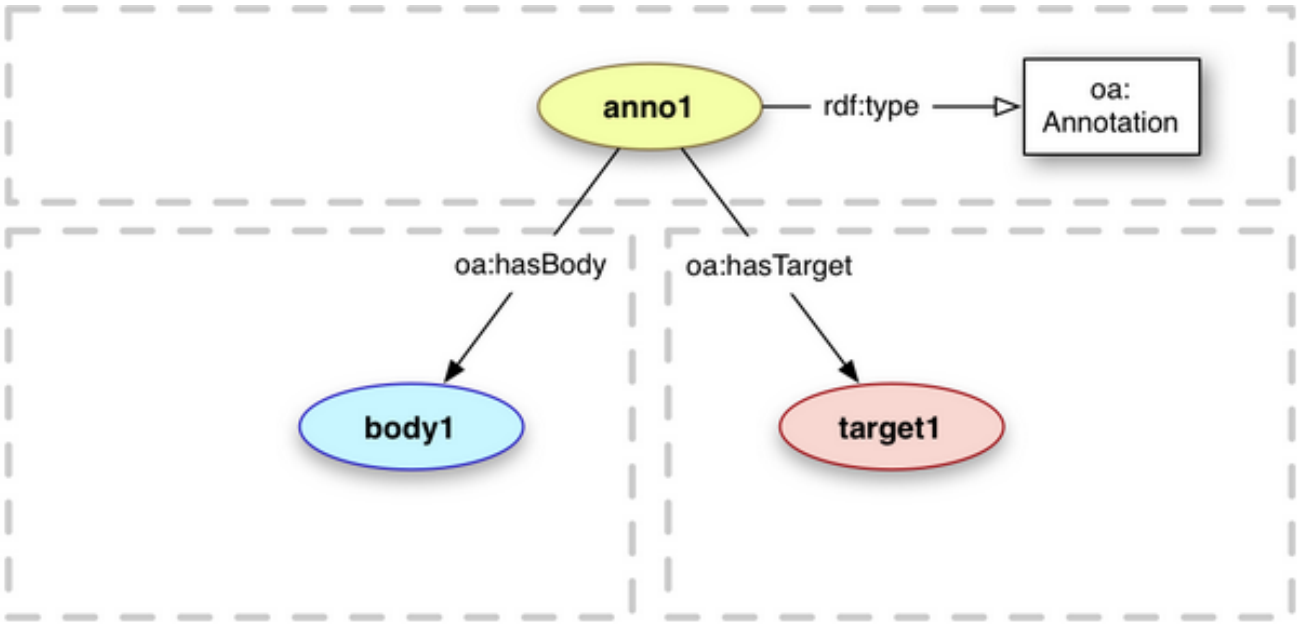


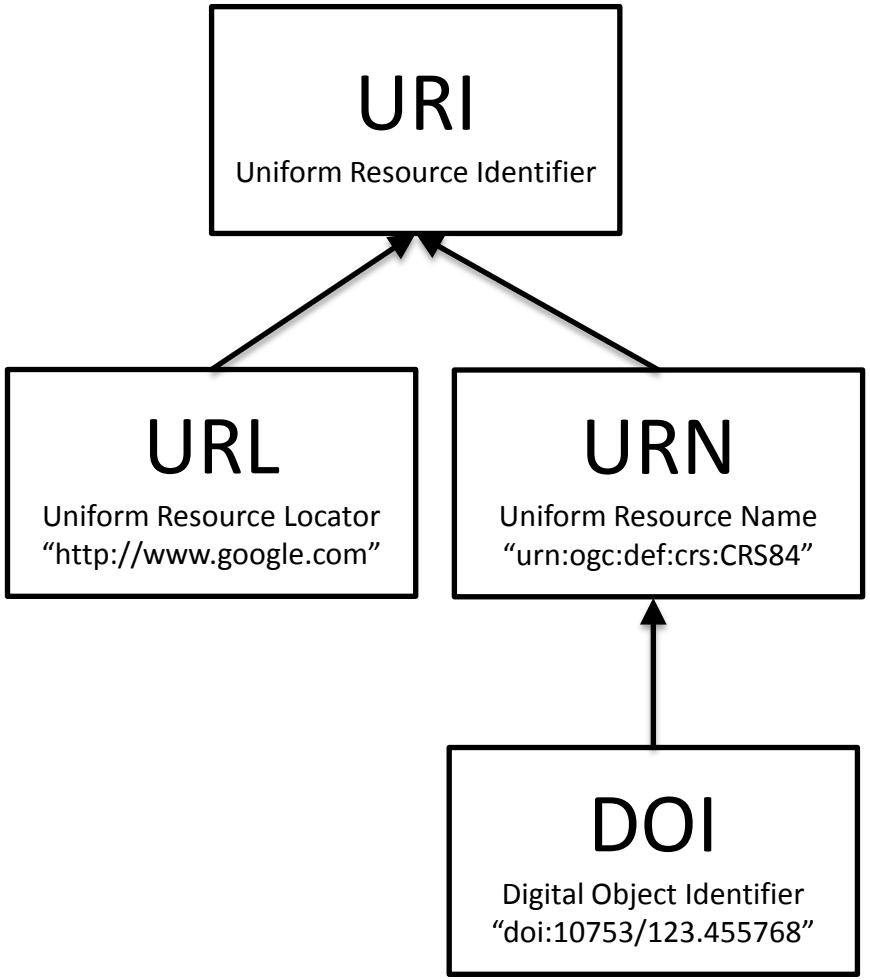
Figure 2.1. Basic Annotation Model

```
<anno1> a oa:Annotation ;
    oa:hasBody <body1> ;
    oa:hasTarget <target1> .
```

http://www.openannotation.org/spec/core/core.html

Important points

- Everything needs a URI!
- "What is a dataset?" is an old chestnut, not yet cracked
  - Means different things in different communities
  - But CHARMe doesn't care: it can annotate anything that has a URI
  - URI hierarchies are managed elsewhere
- Choosing common **vocabularies** is critical
  - Also thesauri, ontologies etc



What CHARMe can enable
(some examples)

Users:

- "Find me all the documents that have been written about this dataset"
  - "… in both peer-reviewed journals and the grey literature"
  - "… and specifically about precipitation in Africa"
  - "… in both STFC's and Astrium's archives"
- "What factors might affect the quality of this dataset?"
  e.g. upstream datasets, external events

Data providers:

- "Who is using my dataset and what are they saying about it?"
- "Let me subscribe to new user comments and reply to them"

What this will not enable

- "Give me the best dataset on sea surface temperature"
- CHARMe will not provide a new "quality stamp" for datasets
  - But will be able to link to such things if other people publish them
- CHARMe will not provide access to actual data
  - (Cf. Web of Science – enables discovery, but access not in scope)
- Not planning to create (another) "one-stop shop" for information
  - We want the information to appear where users are already looking

Some relevant standards

- ISO19156 Observations and Measurements
  - Conceptual model for capturing the information about observations - fundamental to how data is acquired: estimating the value of some property of a feature of interest with a given procedure
  - Includes hooks for associating quality information
- ISO19115 (Quality Package)
  - 'D' (Discovery) Metadata
- ISO19157
  - specifically focuses on quality, improves on and augments ISO19115 Quality Package
- UncertML
  - conceptual model for encapsulating probabilistic uncertainties
- Open Annotation
  - A collaboration focused on an interoperability framework for annotations
  - A data model and ontology
  - Uses Linked Data principles

Some related projects

- GeoViQua
  - Application of ISO19157, integration with UncertML for the capture of uncertainty information
  - Proposed enhancement to ISO19115 aggregation of information for scoping of metadata
- MOLES
  - 'B' (Browse) metadata
  - An application of ISO19156 Observations and Measurements
  - CEDA MOLES implementation
- Metafor CIM 2.0 & ES-Doc
  - Metafor defined a Common Information Model (CIM) to describe climate data and the models that produce it in a standard way
  - ES-Doc expands to generic software and tools for different Earth science data applications
- ESA LTDP (Long-Term Data Preservation)
  - Includes post-fact information e.g. papers
- PREPARDE, OpenAIRE, ORCID, DataCite, OBS4MIPS, EnviLOD … many more!

What have we done so far?

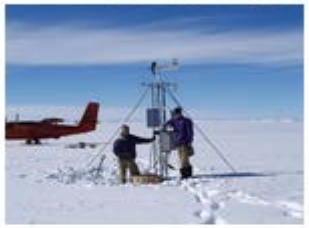
- Collected a set of narrative “user scenarios” from a variety of users
  - Data providers, Data users in various countries
- Currently turning these into formal User Requirements, then into Software Requirements
- Using wireframing and rapid prototyping to help refine requirements
- Made links with related efforts in US and Europe

Can anyone help us with this?

- We would like to find vocabularies/ontologies that:
  - **Describe different kinds of publications** (peer-reviewed journals, technical reports, websites etc);
  - **Describe the relationship between publications and "the things that they are about",** e.g. datasets or sensors.
    - For example, we might want to record that "this publication describes how the dataset was produced", or "this publication reports an issue discovered within the dataset".

Summary / conclusions

- CHARMe will create connected respositories of “commentary metadata”
- Will help users tap into existing expert knowledge about climate datasets
  - But nothing in the project is really specific to climate!
- We will provide this information in existing archives and websites
- Linked Data technologies will enable CHARMe information to be discovered and used in other systems too

CHARMe

SEVENTH FRAMEWORK PROGRAMME

Thank you!

University of Reading

Science & Technology Facilities Council

Royal Netherlands Meteorological Institute
Ministry of Infrastructure and the Environment

CGI

ASTRIUM
AN EADS COMPANY

spotinfoterra
by EADS Astrium Services

Deutscher Wetterdienst
Wetter und Klima aus einer Hand
DWD

Met Office

ECMWF

Jon Blower
j.d.blower@reading.ac.uk
University of Reading
On behalf of all CHARMe partners!

http://www.charme.org.uk