CHARMe: Sharing climate knowledge through commentary metadata and Linked Data (Jan 2013 – Dec 2014)

Victoria Bennett
(Slides thanks to Jon Blower)
on behalf of CHARMe consortium
http://www.charme.org.uk
platform

instrument

algorithm

dataset

scientists

publication

THE ESA CLIMATE CHANGE INITIATIVE
Satellite Data Records for Essential Climate Variables

The ESA Climate Change Initiative is representing and reanalyzing over 40 years of multi-sensor satellite records to generate consistent, transizable, long-term datasets of "essential climate variables" for the climate modeling and research communities.

Published observations from satellite contribute to the Global Climate Observing System (GCOS) network of organizations and satellite data to ensure the consistency, reliability, and interoperability of Earth system observations.
THE ESA CLIMATE CHANGE INITIATIVE

Satellite Data Records for Essential Climate Variables

[Image of satellite and Earth]
How do you fill in the blanks?

• Hope the documentation contains references
• ... to the right versions of things
• ... and using consistent terminology

• In restricted communities, we can gather all information into a central location and harmonize it

• OR we hope that a Google search throws up the right results
Problems

1. Crossing communities is very hard
Problems

1. Crossing communities is very hard

2. Lots of useful stuff isn’t formally documented
Problems

1. Crossing communities is very hard
2. Lots of useful stuff isn’t formally documented
3. Unknown unknowns...
Consequences

• Scientists often use the data that are most easily available, not necessarily the best

• Constant re-discovery of the same issues

• Very hard to share information outside communities
The “Web of Data”
The Web is the “Web of Documents”
Fingers crossed!

• We hope that everything has a document written about it, or we have nothing to cite

• ... and we hope someone writes a new document when “it” changes

• We hope that we are citing the right document (i.e. the most authoritative etc)
“If ... the Web made all the online documents look like one huge book, [Linked Data] will make all the data in the world look like one huge database.”
Why is Linked Data different?

• Focuses on **things**, not documents-about-things

• Requires things to have **unique and persistent identifiers**
  • E.g. DOIs, permanent URLs, etc.

• **Links things together** in a meaningful way

• Links are **readable by computers**
Example: linking a paper to a dataset
(illustrative, not necessarily precise!)

subject predicate object
this is a “triple”
each of these terms has a unique identifier

http://dx.doi.org/12345.678910
http://purl.org/spar/cito/citesAsDataSource
http://www.badc.ac.uk/datasets/sst-cci
The “Web of Data”

THE ESA CLIMATE CHANGE INITIATIVE
Satellite Data Records for Essential Climate Variables

authored

describes / uses / queries ...

carries

produced

produced
CHARMMe:
Sharing climate knowledge through commentary metadata and Linked Data
(Jan 2013 – Dec 2014)
From observations to decisions

Satellites → Observations → Climate Record Creation and Curation → Climate Data Records → Analysis and applications → Reports → Decision making → Decisions, Policies

(Adapted from Dowell et al., 2013), “Strategy Towards an Architecture for Climate Monitoring from Space”
Where can users go for help?

• Scientific literature
  – Huge, verbose and inaccessible to some communities
  – Not well linked to source data

• Technical reports and conference proceedings
  – Hard to find, scattered or inaccessible

• Data centres
  – increasingly strong at providing some important metadata, but don’t usually include community feedback
  – Not all countries and communities have data centres!

• Websites and blogs
  – From CEOS Handbook to a scientist’s blog
  – Increasingly useful, but scattered
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.
How will this be done?

• CHARMe will create connected repositories of commentary information
  – Stored as triples in “CHARMe nodes”,
• Information can be read and entered through websites or Web Services
• Using principles of Open Linked Data
Where does CHARMe fit in?

Supports analysts and scientists in production of information for decision-makers
Challenges

• Ensuring community adoption:
  – We are “injecting” CHARMe capabilities into existing websites used in the community
  – We will “seed” the CHARMe system with information to attract users (e.g. links between publications and datasets)

• Ensuring quality of commentary metadata
  – **Moderation** is a strong user requirement
  – We will provide guides to creators of commentary – what makes a comment helpful?

• Technical challenges:
  • Ensuring that datasets (and missions, instruments...) have identifiers
  • Deciding what a dataset is!
What CHARMe will 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 CEDA’s and Astrium’s archives”

- “What factors might affect the quality of this dataset?”
  e.g. upstream datasets, external events

- “What have other users already discovered about this dataset?”

- “I want to find information related to the dataset I’m looking at”

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 CHARMe will not enable

• “Give me the best dataset on sea surface temperature”
  – The “best” dataset depends on the application

• CHARMe will not provide a new “quality stamp” for datasets
  – But will be able to link to such things if other people publish them, e.g. CORE-CLIMAX Application Performance Matrix, QA4EO certification

• CHARMe will not provide access to actual data
  – (but will help with discovery of data)

• Not planning to create “one-stop shop” for information
  – We want the information to appear where users are already looking
How will CHARMe be used?

- Search for datasets on existing data provider website
- Use the “CHARMe button” to view or record commentary on a particular dataset
- See [http://www.youtube.com/watch?v=F5k_FQt5eb0](http://www.youtube.com/watch?v=F5k_FQt5eb0) for a video of latest prototype
Summary

• CHARMe is using Linked Data techniques to help users of climate data to connect with all the experience in the community
  — Techniques are not specific to climate data

• CHARMe is about the links: other sources provide (most of) the content

• So far we can record documents and free-text annotations about datasets, but can be used to make any kind of links

• Commentary information can be injected into existing websites
Thank you!

victoria.bennett@stfc.ac.uk

http://www.charme.org.uk
“Significant events” viewer

- Google Finance (above) matches stock prices with news events
- CHARMe tool will match climate reanalysis data with “significant events”
  - Algorithm changes, instrument failures, new data sources
- Will allow user annotations on the data and events
- Will be available on the Web
Fine-grained commentary

- Will allow creation and discovery of commentary about specific parts of datasets
- E.g. variables, geographic locations, time ranges
- cf http://maphub.github.io