Understanding climate data through commentary metadata: the CHARMe project

Article

Accepted Version

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Understanding Climate Data through Commentary Metadata: the CHARMe project

J.D. Blower[1,3], R. Alegre[1,3], V.L. Bennett[2,3], D.J. Clifford[1,3], P.J. Kershaw[2,3], B.N. Lawrence[1,2,4], J.P. Lewis[1,3], K. Marsh[2,3,4], M. Nagni[2], A. O'Neill[1,3], R.A. Phipps[1,3]

[1] Department of Meteorology, University of Reading, Reading, United Kingdom
[2] STFC Centre for Environmental Data Archival, Rutherford Appleton Laboratory, Didcot, United Kingdom
[3] National Centre for Earth Observation, United Kingdom
[4] National Centre for Atmospheric Science, United Kingdom
j.d.blower@reading.ac.uk

Abstract. We describe the CHARMe project, which aims to link climate datasets with publications, user feedback and other items of "commentary metadata". The system will help users learn from previous community experience and select datasets that best suit their needs, as well as providing direct traceability between conclusions and the data that supported them. The project applies the principles of Linked Data and adopts the Open Annotation standard to record and publish commentary information. CHARMe contributes to the emerging landscape of "climate services", which will provide climate data and information to influence policy and decision-making. Although the project focuses on climate science, the technologies and concepts are very general and could be applied to other fields.

Keywords: Climate Data, Climate Services, Open Annotation, Linked Data, Semantic Web, metadata

1 Introduction and motivation

Accurate, long-term monitoring of the Earth is of vital importance for gathering information about our climate. This information, in turn, forms an important part of the evidence base for operational and policy decisions that have far-reaching effects on society. Climate data are used by both the public and private sectors for applications such as controlling greenhouse gas emissions, energy production, food security and flood prediction [1].

Climate data come from various sources and encompass many types of physical, chemical and biological variables. Networks of *in situ* sensors have been used in climate studies for over 75 years and provide (generally) highly-accurate measurements at a limited set of points in space and time. In the past few decades, the importance of space-based measurements from satellites (i.e. *Earth Observation*) has increased. Satellites provide measurements of a highly diverse range

of variables, with the key advantage that measurements can be produced at a range of spatial scales (local, regional and global). Further information about our climate comes from numerical simulations, which capture (as best we can) our knowledge of fundamental physical, chemical and biological processes in order to make predictions of the future. These different systems have complementary strengths and weaknesses, therefore many problems in modern environmental science (including weather forecasting for example) are tackled by combining multiple sources of information.

Users of climate data are highly diverse, ranging from research scientists (for example, searching for signals of long-term climate change) through government policy-makers (for example, setting caps on carbon dioxide emissions) to operational decision-makers (for example, planning construction of flood defences). All these users require access to expert knowledge to help them to decide which climate datasets to use in their studies and to understand the fitness for purpose of those datasets for their problem.

The climate science community has made great strides in the past decade in providing high-quality metadata to help users to discover and use climate data (e.g. [2]). Most of this metadata encompasses information about the *intrinsic* characteristics of datasets (e.g. spatial and temporal resolution and coverage), although increasingly information is being released also about dataset provenance (i.e. the processes that led to the production of the dataset). One important aspect that is currently very little addressed is the systematic publication of information about how a dataset has been used by the community. This information is useful for several reasons, including:

1. Usage information helps new users to select between apparently similar datasets to choose the best dataset for their purpose, in a similar manner to the use of reviews on a shopping or travel website.
2. It increases the probability that vital results and lessons concerning the strengths and weaknesses of datasets are retained by the community, helping to avoid reinvention and the loss of information that is caused when results are not formally published (or are published in locations that are obscure to the user in question).
3. It provides another view of *data quality* (in the sense of "fitness for purpose") to complement other quality information that should be reported by the data provider (such as accuracy, uncertainty and metrological traceability).
4. It increases the traceability of conclusions back to their source data and the reproducibility of results (e.g. the draft 3rd US National Climate Assessment [3]), refers to the importance of the "line of sight between conclusions and data").
5. It provides a new route to data discovery, particularly where users record information about how datasets relate to each other.
6. It provides valuable feedback to data providers, as it helps them to improve their data and report back to their own funding agencies.

Publication of such usage information is currently difficult, primarily because the metadata/information paradigm around environmental data is mostly

provider focused, and where third party (user) information exists, it is not easily discoverable alongside the original data. Here we describe the "CHARacterization of Metadata to enable high-quality climate applications and services"(CHARMe) project, which aims to provide mechanisms to address this problem, applied to some specific problems in earth observation.

In the remainder of the paper, we first introduce an information paradigm for discussing metadata and the concept of "commentary metadata". We then describe the CHARMe project itself, the expected users and their requirements in the context of our information paradigm, before introducing the CHARMe technical approach. The paper concludes with a summary and brief discussion of future work.

2 Commentary Metadata and CHARMe

2.1 A short introduction to metadata

Lawrence et al [4] provide a taxonomy of different kinds of metadata employed in data infrastructures:

- Archive (A) metadata encompasses precise descriptions of information such as spatial and temporal referencing at the level of individual samples. A-metadata is required to actually use the data in calculations.
- Browse (B) metadata supports understanding the context of data (including provenance) and choosing between similar datasets, and generally conforms to some community standard metadata semantics.
- Commentary (C) metadata includes citations of the data and post-fact assertions about quality. In their original taxonomy, Lawrence et al use the term "Character" metadata in their paper, but we feel that the word "Commentary" is more helpful here, particularly since we expand the scope of C-metadata to go well beyond those envisaged in the original taxonomy.
- Discovery (D) metadata is typically harvested into catalogues, encompassing summary information about dataset contents and overal spatial-temporal extent.
- Extra (E) metadata is highly discipline-specific (e.g. highly structured descriptions of sensors or documents describing experimental protocols).

The boundaries between these types are not completely sharp, but, in general, Archive, Browse and Discovery and Extra metadata are intrinsic to the dataset and hence known to the data provider *a priori*. Commentary metadata (hereafter referred to as "C-metadata") are normally produced *after* the dataset has been published and reflects real use in the community. C-metadata will typically therefore be closely allied with the particular application to which the dataset has been put, and is therefore *extrinsic* to the dataset itself.

Examples of C-metadata include peer-reviewed publications, technical reports, third-party quality assessments and error characterizations, together with more informal material such as websites, blog entries and ad-hoc comments.

These kinds of C-metadata are the focus of this paper. However, we know from interactions with users that other kinds of information, which might formally be thought of as B- or E-metadata (such as properties of data distribution or service interface descriptions), are often made available by third parties. So here we include within our definition of C-metadata, intrinsic data properties which typically have not yet been published by the data providers alongside their data. (Such metadata might even be published by third parties, but it might even initially be published using the C-metadata formalism by the data providers themselves, before formal ingestion into their B- or E-metadata systems). Additional important categories of information include relevant events such as large volcanic eruptions and instrument failure logs. Although many of these are not discussed in detail in this paper they will nevertheless be addressed in the CHARMe project (see section 5).

One way of thinking about these metadata categories is that the intrinsic metadata (A,B and E) form part of a provider data infrastructure, and C-metadata effectively describes the wider extended information ecosystem around data objects. A,B and E are generally managed, C evolves.

The importance of all such metadata is well understood — being ranked as important as instrinsic data quality [5]. However, it's also known that acquiring quality metadata is difficult, for example:"It's fine to say that scientists should record and preserve all this information, but it is far too laborious and expensive to document everything. The scientist wants to do science, not be a clerk." [6]. One method of addressing this is to introduce systems which reward quality metadata production, such as data publication (e.g. [7]), another is to make it easier to establish and use an information ecosystem (e.g. [8]). We think that both approaches are important, but here we concentrate on supporting such an information ecosystem, built around the concept of commentary metadata. However, we note that from a user perspective these notions of distinguishability between data infrastructure, external ecosystems and data Publication, should be lost in a broad spectrum of information tools with which they interact.
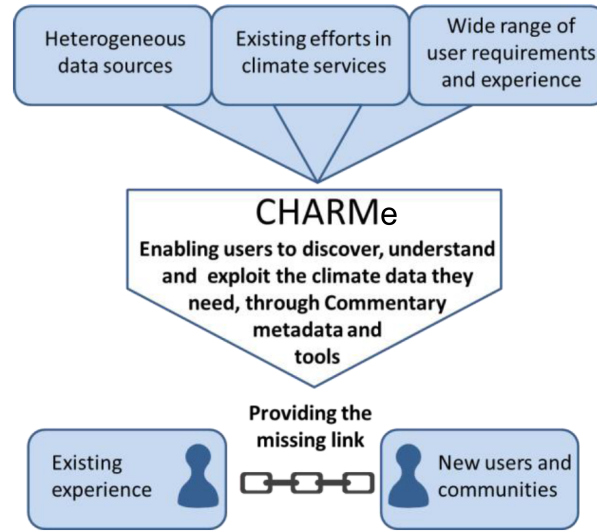
2.2   The CHARMe project

Although the notion of C-metadata is quite general and could be applied in any scientific discipline, the purpose of the CHARMe project is to apply the concept specifically to the use of satellite-derived climate data. The CHARMe project aims to connect users, and prospective users, of climate datasets with the previous expertise that has accumulated in the community, and enable them to contribute back information on their own experiences.

The main objectives are:

1. To develop an open-source system for recording "Commentary metadata" that links with climate datasets and other sources of information such as descriptions of sensors and instruments.
2. To provide interfaces for Commentary metadata to be entered, queried and displayed through existing community websites and machine-readable interfaces (section 4).

3. To identify, and engage with, key strategic stakeholders (including climate data users, producers and high-level global initiatives) to ensure that the CHARMe concept is understood and supported by providers and users of climate data.

4. To develop tools that demonstrate other ways in which Commentary metadata can be produced and exploited in a variety of scenarios (section 5).

Fig. 1. Overview of the aims of the CHARMe project.

The CHARMe project operates in a wider context of the development of "Climate Services". Climate Services are conceived to be activities that produce climate-related information for policy-makers and decision-makers, in order to benefit society. Satellite-derived information will form an important component of future climate services, and it is recognized that there is a need for satellite data to be curated and shared in a systematic manner [9]. Fig. 2 sketches a logical view of the transformations from remote sensing measurements to decisions. Data will be processed and re-processed multiple times, leading to a complex landscape of interrelated datasets, which require strong expertise (and good quality metadata) to interpret correctly. The Global Climate Observing System (GCOS) has published guidelines for data producers [10], to ensure good quality documentation for climate records, including a recommendation for data providers to provide a user feedback facility (guideline 10).

In the context of global climate services, therefore, CHARMe aims to help record, retain and disseminate community experience with interpreting climate data, and provide a means for feeding back this experience to the data providers.
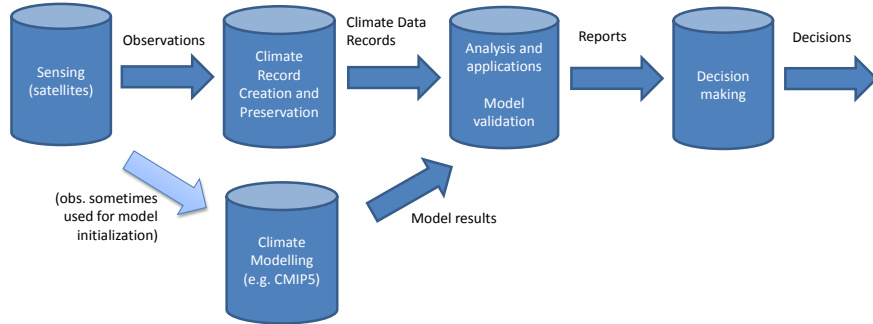
Fig. 2. The main components of a logical architecture for climate services, adapted from Dowell et al (2013) [9]. We have added an extra activity to Dowell et al's diagram to encompass numerical modelling activities. Such models sometimes use satellite observations to set up their initial state. Analysts intercompare results from different models (e.g. in the Coupled Model Intercomparison Projects) and validate results, where possible, against observations. Both models and observations feed into reports (such as the IPCC Assessment Reports) that are used in decision-making and policy-making.

3 Users and requirements

3.1 Who are the users?

The CHARMe project has considered the needs of several kinds of end-user stakeholder, including scientists/researchers, commercial users and policy-makers. (In addition the project considers other kinds of stakeholder including data providers, auditors, quality assurance professionals and system administrators, although these are not discussed in detail in this paper.)

Although a key aim of CHARMe is to provide information that influences policy-making, Fig. 2 above shows that policy-makers will probably not usually be *direct* users of the CHARMe system. Instead, groups of analysts and scientists will use Climate Data Records directly, producing reports, digests and assessments for the policy-makers. CHARMe is therefore seeking to interact most closely with analysts (in academia, private consultancies, humanitarian institutions, government departments and elsewhere) who are tasked with the production of information for informing policy-making. The goal of CHARMe is to help these people select and use the best information available to them, and to provide guidance on interpreting the data correctly.

3.2 What do they need?

CHARMe has performed an initial stage of requirements-gathering in order to inform the general technical approach described in section 4 below. A full discussion of this process is beyond the scope of this paper, but a summary is

presented here. We engaged a variety of users to assess (i) what they use climate data for, (ii) where they obtain climate data from, (iii) how they judge a dataset's fitness for their purpose, (iv) what information needed to make that judgement is currently hard to find, and (v) what tools they think would increase their understanding of climate data. In addition to these structured questions, we also allowed the users to respond in free text, in order to describe their scenarios in more detail. So far we have received 42 responses from scientists and modellers, commercial users, humanitarian institutions and policy-makers. These cover a wide range of scenarios including investigations of climate change trends, seasonal forecasting, validation of ocean reanalyses, mapping of natural resources, monitoring of extreme events and agricultural impact studies. In addition, we held a small user workshop. Attendees represented a wide range of backgrounds, including academic researchers, Earth Observation consultants in industry, government agencies (including the UK Space Agency) and providers of meteorological services. These potential users and CHARMe consortium participants exchanged ideas through brainstorming sessions, resulting in some new ideas and requirements, discussed below. Additional user workshops will be held in future stages of the project to better assess CHARMe's focus.

The main findings from these interactions were:

1. C-metadata is important and a currently-unmet need. However, users require all the other kinds of metadata too (see section 2.1) and so the metadata should be presented in an integrated fashion.
2. There was a great deal of commonality in the types of C-metadata that were requested. The most important of these were publications, including both peer-reviewed publications and other items such as technical reports (these can be currently hard to find because they are not usually systematically catalogued and lack a dedicated search engine). Other important items were user feedback, software support, assessments of errors and quality, maturity indices (e.g. [11]) and traceability information.
3. The need for high quality of the commentary itself, which includes such concerns as editing, de-duplication and, crucially, moderation. Users are concerned that other users could, maliciously or unintentionally, severely compromise the adoption of a dataset by making unfair or inaccurate comments. Thus, we have incorporated requirements on registration and authentication.
4. Other suggestions included a tool for "subscribing" to comments about a dataset (so that new information is automatically sent to the user) and supporting discussions (i.e. making comments on comments).

A feature that was frequently requested by users is to have a "quality stamp" that highlights the "best" dataset out of a choice of several. After careful consideration, we have rejected this requirement as out of scope for CHARMe as we believe that the choice of the "best" dataset is a highly subjective one that depends strongly on the application. Other groups have reached similar conclusions: for example, current developments to establish a "GEO Label" do not focus on the subjective quality of the dataset itself, but an objective summary

of the presence or absence of key metadata items [12]. CHARMe may follow a similar path.

4 Technical approach

4.1 Data models for C-metadata

As discussed in section 2.1, C-metadata is different from other kinds of metadata. It is driven primarily by users of data (not producers), it can accumulate and change rapidly and it is highly diverse in nature, linking together information from many different sources. This presents several challenges when trying to apply traditional standards-based data models (e.g. ISO and OGC) and traditional model-driven development approaches, since the field is too immature and dynamic to enable the robust definition of a fixed data model *a priori* that can be sustained.

We plan to follow an iterative development methodology and we therefore sought a flexible, overarching data model that encompasses the general notion of commentary, permitting further sophistication to be added gradually as our developments and experiments progress. The recently-published Open Annotation (OA) standard from the World Wide Web Consortium (W3C) provides us with a very attractive solution, since the concept of an "annotation" is very similar to our notion of commentary metadata. OA is based on Linked Data principles and defines a simple and general data model, based on the Resource Description Framework (RDF) for recording annotations about objects. An annotation associates a piece of information (the *body*) with a subject (the *target*). This maps directly to CHARMe requirements: a target could be a dataset and an annotation body some user comment about that dataset, or a reference to a publication. OA supports the concept of a *motivation* for an annotation. This is a controlled list of terms, including *linking*, *describing*, *bookmarking* and *questioning*. These terms, together with terms from other vocabularies such as CiTO [13] can provide valuable semantic information that describes more precisely the relationship between publications and datasets (see Fig. 3).

Deliberately, no types are set for the target or body: users of OA can import terms from other vocabularies and ontologies to record types of interest. This enables the body of one annotation to be the target of another allowing for example, the creation of a chain of annotations in the manner of an online discussion forum. Annotations can contain multiple targets and bodies, enabling, for example, the linking of a publication to more than one dataset, a key CHARMe requirement.

As with all approaches based on Linked Data principles, the use of globally-unique and persistent identifiers is extremely important. A system for identifying formal peer-reviewed publications (using Digital Object Identifiers, DOIs) is now well-established, and the use of DOIs in identifying datasets is becoming more established. Web references (e.g. URLs to websites and blog entries) can be used, although they provide no guarantee of persistence.

```
@prefix oa:      <http://www.openannotation.org/spec/core/> .
@prefix dctypes: <http://purl.org/dc/dcmitype/> .
@prefix cito:    <http://purl.org/spar/cito/> .

<anno1> a oa:Annotation ;
    oa:hasBody     <citation> ;
    oa:hasTarget   <http://dataprovider.org/datasets/sst> ;
    oa:motivatedBy oa:linking .

<citation> a cito:CitationAct ;
    cito:hasCitingEntity  <http://dx.doi.org/12345.678910> ;
    cito:hasCitationEvent cito:citesAsDataSource ;
    cito:hasCitedEntity   <http://dataprovider.org/datasets/sst> .

<http://dx.doi.org/12345.678910> a dctypes:Document .
<http://dataprovider.org/datasets/sst> a dctypes:Dataset .
```

Fig. 3. Example of modelling the link between a publication and a dataset using the Open Annotation (OA) data model together with the CiTO ontology [13]. The citation records that the dataset was cited as a data source within the publication in question; other CiTO terms could be used to record different relationships, for example that the publication *describes* the dataset. The example could be extended to associate the publication with multiple datasets, perhaps with different types of relationship for each link.

We may compare this general and flexible approach with the more prescribed approach taken by the user feedback model of the GeoViQua system [14], which also aims to record information about dataset usage. In GeoViQua, the key class is a *FeedbackItem*, which defines a fixed set of fields that encompass user comments, citations and the identity of the commenter. This approach has the advantage that feedback items have a consistent and structured form that can be entered and queried in a defined manner. However, if a user of the schema wishes to add new fields or employ different semantics (for example, to record a publication that references several datasets in different ways) this cannot easily be done without defining a new information class. By contrast, the OA approach allows users to express their intent however they wish, but querying and aggregating annotations becomes more difficult, unless there is wide agreement upon the use of particular structures and terms (this is a general challenge of Linked Data and the Semantic Web). Therefore, although OA provides a useful "top level" data model, CHARMe will need to develop more structured data models to encode the bodies of annotations, to enable querying and comparison.

The two approaches are not mutually exclusive and can be at least partly compatible and complementary. It is possible to express a GeoViQua Feedback-Item using the OA model, although the converse is more difficult because the OA model is more general. Practical experience will yield more information about their relative strengths and weaknesses.

4.2 Architecture of a CHARMe node

A conventional client-server architectural pattern is envisaged for CHARMe with C-metadata stored in a repository served over web service interfaces to expose it to various client applications enabling them to search, add to, modify and delete metadata entries. Access control will need to be implemented for the latter three operations in order for comments to be attributable to their creators, and to ensure that modification and moderation are controlled.

Support for multiple "CHARMe nodes" (i.e. instances of the repository and services) is envisaged. This will enable individual organisations to host their own service instances. These could be linked in a federated model such that metadata can be shared or replicated between a number of trusted nodes. Alternatively some organisations may wish to host their own private node within an intranet.

The web service interfaces are envisaged to expose both SPARQL and RESTful query models (Fig. 4). Setting a standard interface to clients will make it possible to easily interchange the implementation of the persistence layer. Apache Jena has been used for initial tests. This uses the combination of a triple store for storing RDF together with Apache Lucene or Solr to create an index for free text search.
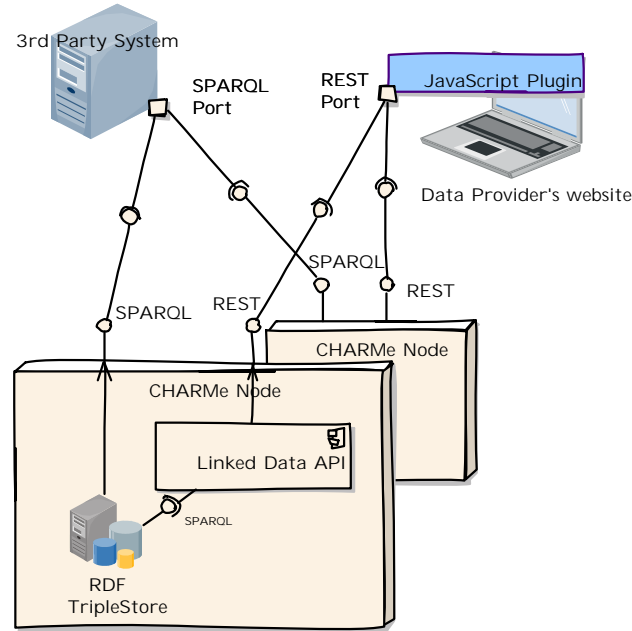
Fig. 4. Outline architecture of a CHARMe node. Annotations are stored in a triple-store, exposed through a SPARQL interface for querying. Using a translation layer (e.g. the Linked Data API, https://code.google.com/p/linked-data-api/) a RESTful interface is also provided, which provides a more convenient querying interface for many types of client.

4.3 User interface

A strong user requirement is the need to draw the various pieces of metadata about a given dataset together into one place, whether they come from the original data provider or the CHARMe system. We also wish to ensure that C-metadata is available in the locations from which users already obtain climate data. Therefore, instead of creating a new web portal, we plan to develop plugins for existing data-access portals (see Fig. 4). This follows the approach of the Metafor project's metadata viewer (http://metaforclimate.eu), which uses a JavaScript plugin to inject new metadata into existing websites that serve data from climate simulations. There are some practical concerns with this approach (including security), which are beyond the scope of this paper but will be carefully considered in the CHARMe project.

Irrespective of the technology approach, the design of the user interface presents a strong challenge. The Open Annotation data model allows flexibility at the level of the database of C-metadata, but the user interfaces must collapse all the abstract possibilities afforded by RDF into particular concrete structures (menu items, forms etc). This is another reason to constrain the bodies of annotations into a more formal data model. We plan to mitigate these concerns by adopting a modular approach to the user interface design and by closely coupling individual interfaces to the underlying C-metadata types they expose.

5 Summary and future work

We have described the motivation behind, and early progress of, the CHARMe project, which aims to assist users and providers of climate data by sharing experience through "commentary metadata". A primary use case is linking climate datasets with publications that use, describe or evaluate them. Our proposed approach, centred on Linked Data and the Open Annotation standard, will give flexibility and interoperability, but more investigation is required in certain areas, chiefly the design of appropriate and useful user interfaces.

In addition to the developments described above, we also plan to explore more advanced methods for creating and using commentary metadata. These include the development of a faceted search facility, an intercomparison tool for data and metadata and a mechanism for associating C-metadata with *subsets* of datasets, such as spatial regions or particular points in time. This will enable "fine-grained commentary" and will allow users to correlate "significant events" (such as volcanic eruptions or satellite instrument failures) with features in the datasets.

Although the CHARMe project focuses on climate science, the approach is very general and could be applied readily to other fields.

Acknowledgements

References

1. Core Writing Team, R.K. Pachauri, A. Reisinger: IPCC, 2007: Climate change 2007: Synthesis report. contribution of working groups I, II and III to the fourth assessment report of the intergovernmental panel on climate change. Technical report, Geneva, Switzerland (2008)
2. Lawrence, B.N., Balaji, V., Bentley, P., Callaghan, S., DeLuca, C., Denvil, S., Devine, G., Elkington, M., Ford, R.W., Guilyardi, E.: Describing Earth System simulations with the Metafor CIM. Geoscientific Model Development **5** (2012) 1669–1689 doi:10.5194/gmdd-5-1669-2012
3. National Climate Assessment and Development Advisory Committee: 3rd national climate assessment (draft). http://ncadac.globalchange.gov/ (2013)
4. Lawrence, B., Lowry, R., Miller, P., Snaith, H., Woolf, A.: Information in environmental data grids. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **367** (March 2009) 1003–1014 doi:10.1098/rsta.2008.0237
5. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst. **12**(4) (March 1996) 5–33
6. Gray, J., Szalay, A.S., Thakar, A.R., Stoughton, C.: Online scientific data curation, publication, and archiving. In: Astronomical Telescopes and Instrumentation. (2002) 103–107
7. Lawrence, B., Pepler, S., Jones, C., Matthews, B., Callaghan, S.: Citation and peer review of data: Moving towards formal data publication. International Journal of Digital Curation **6**(2) (2011) 4–37 doi:10.2218/ijdc.v6i2.205
8. Parsons, M.A., Fox, P.A.: Is data publication the right metaphor? Data Science Journal **12** (2013) WDS32–WDS46 doi:10.2481/dsj.wds-042
9. Dowell, M., Lecomte, P., Husband, R., Schulz, J., Mohr, T., Tahara, Y., Eckman, R., Lindstrom, E., Wooldridge, C., Hilding, S., J.Bates, Ryan, B., Lafeuille, J., Bojinski, S.: Strategy towards an architecture for climate monitoring from space. http://www.wmo.int/pages/prog/sat/documents/ARCH_strategy-climate-architecture-space.pdf (2013)
10. World Meteorological Organization: Guideline for the generation of datasets and products meeting GCOS requirements (GCOS-143). https://www.wmo.int/pages/prog/gcos/Publications/gcos-143.pdf (2010)
11. Bates, J.J., Barkstrom, B.: A maturity model for satellite-derived climate data records. In: 14th Conference on Satellite Meteorology and Oceanography, Atlanta, USA. (2006)
12. Lush, V., Bastin, L., Lumsden, J.: Developing a GEO Label: Providing the GIS community with quality metadata visualisation tools. In: Proceedings of the 21st GIS Research UK (GISRUK 3013), Liverpool, UK, 3 - 5 April 2013. (2013)
13. Shotton, D., Peroni, S.: CiTO, the Citation Typing Ontology. http://purl.org/spar/cito/
14. Yang, X., Blower, J.D., Bastin, L., Lush, V., Zabala, A., Masó, J., Cornford, D., Díaz, P., Lumsden, J.: An integrated view of data quality in earth observation. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **371**(1983) (2013) 20120072 doi:10.1098/rsta.2012.0072