Capturing and sharing our collective expertise on climate data: the CHARMe project

Capturing and Sharing Our Collective Expertise on Climate Data

The CHARMe Project

BY DEBBIE CLIFFORD, RAQUEL ALEGRE, VICTORIA BENNETT, JON BLOWER, CECELIA DELUCA, PHILIP KERSHAW, CHRISTOPHER LYNNES, CHRIS MATTMANN, RHONA PHIPPS, AND IRYNA ROZUM

MOTIVATION. Increasingly, people gather inputs and make decisions through websites that offer evolving commentary from others at the point of decision or purchase; think of Amazon, TripAdvisor, Yelp, and countless others. Introducing an analogous capability into Earth science data selection and acquisition has the potential to turn what is currently a solitary exploration into one where the user's decisions are informed by a broad community.

Today, science users in search of relevant datasets for their investigations find these data in a variety of ways: filtering by climate parameter, crawling/browsing data servers, or through some other sophisticated means. Regardless of how a user arrived there, ultimately she is presented with a list of links to data sources (files) through some data system interface. These are what the science user accesses to compute, analyze, and visualize information, yet many times they lack up-to-date ancillary information—for example, pointers to documentation on the dataset, contact information for the dataset engineer or the responsible science lead, information on how others have used the data, or if there are known problems. In all of these cases the interested user must navigate away from the search results in order to discover the information they require, and it is not always obvious how to go about it, or whether the information she discovers is valid for, or even relevant to, the particular dataset of interest.

The European collaborative project CHARMe (Characterization of metadata to enable high-quality climate services) has developed a system that avoids these navigation and presentation issues by providing crowd-sourced user commentary directly next to the data download link. The connection is immediate and obvious, and content continues to expand as the data get more use. Users can post notes about issues and questions to the data providers and other users, and the data engineers responsible for the dataset know exactly what data are involved when answering questions. The technology allows for this knowledge capture to stay connected to the dataset itself, no matter how the user arrives at the download link.

As an example, Fig. 1 shows CHARMe being used via the website of the Climate Monitoring Satellite Applications Facility (CMSAF), hosted by Deutscher Wetterdienst: www.cmsaf.eu/doi. In this case the user is browsing a list of datasets, and having clicked on the blue "C" icon is viewing an annotation linking the selected dataset to a validation report. The tags "describing" and "linking" (as well as others available to the user when submitting the annotation) help subsequent users to understand why the comment was made, and discover the comment using a facility known as a "faceted search."

This article describes the application of this system to climate data, the underlying technology and data model, and the tools that have been developed to demonstrate use of this commentary to explore climate data in new ways.

AFFILIATIONS: CLIFFORD, BLOWER, AND PHIPPS —University of Reading, Reading, United Kingdom; ALEGRE—University College London, London, United Kingdom; BENNETT AND KERSHAW—Science and Technology Facilities Council, Swindon, United Kingdom; DELUCA—NESII/CIRES/NOAA Earth System Research Laboratory, Boulder, Colorado; LYNNES—NASA Goddard Space Flight Center, Greenbelt, Maryland; MATTMANN—NASA Jet Propulsion Laboratory, Pasadena, California; ROZUM—European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom
CORRESPONDING AUTHOR: Debbie Clifford, Department of Meteorology, University of Reading, Earley Gate, Whiteknights, Reading, UK RG6 6BB
E-mail: d.j.clifford@reading.ac.uk

DOI:10.1175/BAMS-D-14-00189.1

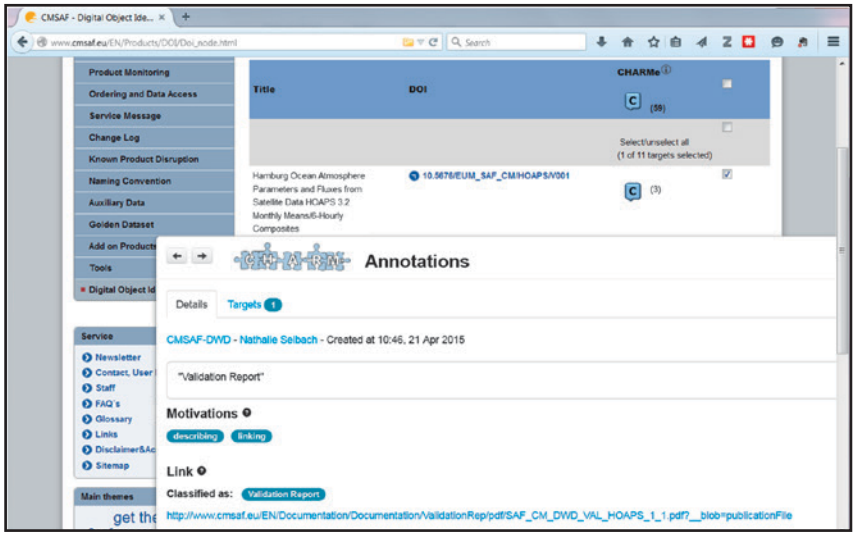


Fig. 1. An example of an annotation that describes a dataset found via the CMSAF website.

notes that describe the characteristics of the observational data in a way that is tailored to the needs of climate modelers.

The Climate Data Guide from the National Center for Atmospheric Research (https://climatedataguide.ucar.edu/) allows users to compare the attributes, strengths, and limitations of multiple datasets. However the website specifically states "The Climate Data Guide generally does not distribute data sets. *It is your responsibility to find and process the data that you need*." Again, the data and supporting information are in different locations, and it is up to the user to navigate between them.

LINKING COMMENTARY TO CLIMATE DATA. Climate data are diverse, encompassing in situ and remotely sensed observations, the results of numerical models, and the combination of models and observations in reanalysis programs. A particular feature of climate data is that their intrinsic value grows over time, but there is a risk that the expertise in the use of the data is lost as people move on and expert teams disperse. End products are often derived from a variety of sources, making it difficult to issue simple statements about a product's quality, and impossible to label a particular dataset as "the best." Instead, users need to weigh up a range of features to judge a dataset's fitness for their specific purpose. The importance of this "knowledge around the data" is recognized by the International Strategy Toward an Architecture Climate Monitoring from Space (Dowell et al. 2013): it is just as important to preserve this knowledge about the data as it is to preserve the measurements themselves.

There are many international collaborations and initiatives already gathering the information users need about climate data. For instance, the European project "Coordinating Earth Observation data for reanalysis for climate services: CORE-CLIMAX" (www.coreclimax.eu/) is bringing together the data and information to support reanalyses of past climate. The international Obs4MIPS (Observations for Model Intercomparisons; www.earthsystemcog.org/projects/obs4mips/) activity is making observational products more accessible for climate model intercomparisons, partly through the generation of technical

In addition, every Earth science researcher and climate data user around the world will be generating commentary as they go about their work. It is surely the case that the same advances and dead-ends are being discovered time and again. Traditionally, such knowledge is captured in narrative form and shared through human-readable means such as papers, articles, and presentations; a great deal of extra value can be gained by sharing this information in a machine-readable, searchable way.

The term "Linked Data" refers to a set of best-practice techniques that describe how one can make data available on the web and interconnect it with other data, with the aim of increasing its value for

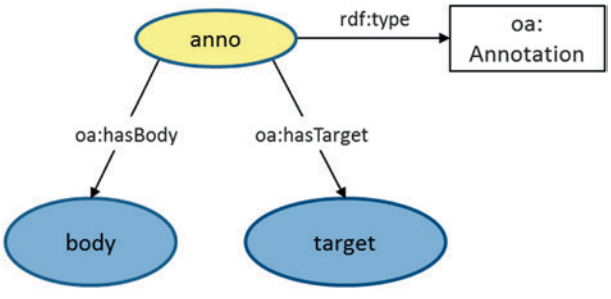


Fig. 2. The basic Open Annotation (OA) data model showing how an annotation links one piece of information (held in the body) with another (the target), and is described using the Resource Description Framework (RDF).

applications and users. The CHARMe project applies the principles of Linked Data to climate data commentary: the representation of the commentary within a formal data model is a critical part of the CHARMe design. Open Annotation, a W3C effort to develop a common approach to annotating digital resources, provides the underlying concept (www.openannotation.org/). It offers a simple and general data model for recording annotations about objects. An annotation associates a piece of information (the body) with a subject (the target), as shown in Fig. 2. Although applied largely to arts- and humanities-related applications thus far, the model has shown itself to be versatile and readily adaptable to Earth observation and climate science use. For example, Open Annotation provides a means of specifying subsets of a given target, such as a character range to reference a given piece of text from a document. Building on these concepts in the framework, it is possible to define extensions to describe *geographic* subsets of datasets, which are described further in the section below, "Exploring Commentary in Space and Time."

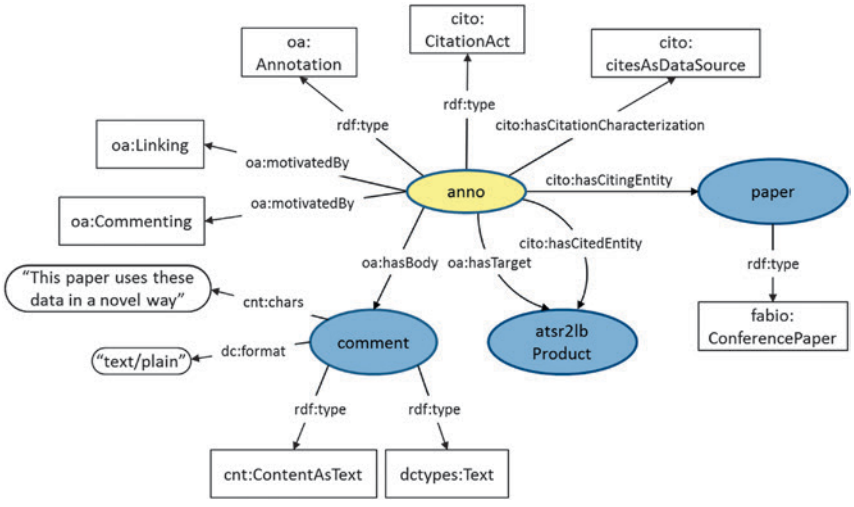


Fig. 3. The CHARMe data model being used to link a comment about a conference paper to the data it cites (in this example, the "ATSR2lb product"). The links are encoded using standard ontologies for describing resources such as RDF and Dublin Core (dc); see the For Further Reading section for more details.

In a data portal, a target is typically a dataset (or subset of a dataset), while the annotation body holds the commentary. The overall design offers flexibility: a single comment body can be associated with many data targets, or the body of one annotation can be the target of another. This is illustrated in Fig. 3, which shows how the CHARMe data model represents a comment on a conference paper, also capturing a link to the dataset cited by the paper. In this way the CHARMe system begins to link the user to a "web of knowledge" about the data they are interested in.

The GeoViQua project (www.geoviqua.org/) has developed a data model for user feedback on datasets in the Global Earth Observation System of Systems. This model shares some conceptual similarities with the CHARMe data model, with the main difference being that the CHARMe model is built on Linked Data principles and the RDF (Resource Description Framework; www.w3.org/RDF/) data model, whereas the GeoViQua model is based around a UML (Unified Modeling Language; www.uml.org/) model and a derived XML encoding. These two approaches are complementary. UML models describe information in a relatively fixed, rigid fashion; this allows data producers and consumers to interoperate closely because the consumer knows exactly what data structure to expect. The disadvantage of this is that data models, once fixed and agreed upon, can be hard to apply to situations that were not expected at design time. By contrast, RDF models enable the data producer to structure data more flexibly, enabling new requirements to be more smoothly integrated. However, it can be difficult to write data-consuming software that can handle all the possibilities afforded by this high degree of flexibility. Discussions are ongoing within the Open Geospatial Consortium to harmonize the CHARMe and GeoViQua models at the conceptual level, enabling implementers to apply the encoding they feel is most appropriate to the application.

TECHNICAL ARCHITECTURE. Rather than create a new web portal to expose climate data commentary to users, CHARMe has developed a plug-in that is simple to include in existing data-access portals. The plug-in highlights to users the existence of commentary on their datasets of interest and allows them to make comments of their own. A third-party data provider is "CHARMe enabled" by integrating the JavaScript for the plug-in in their website. As

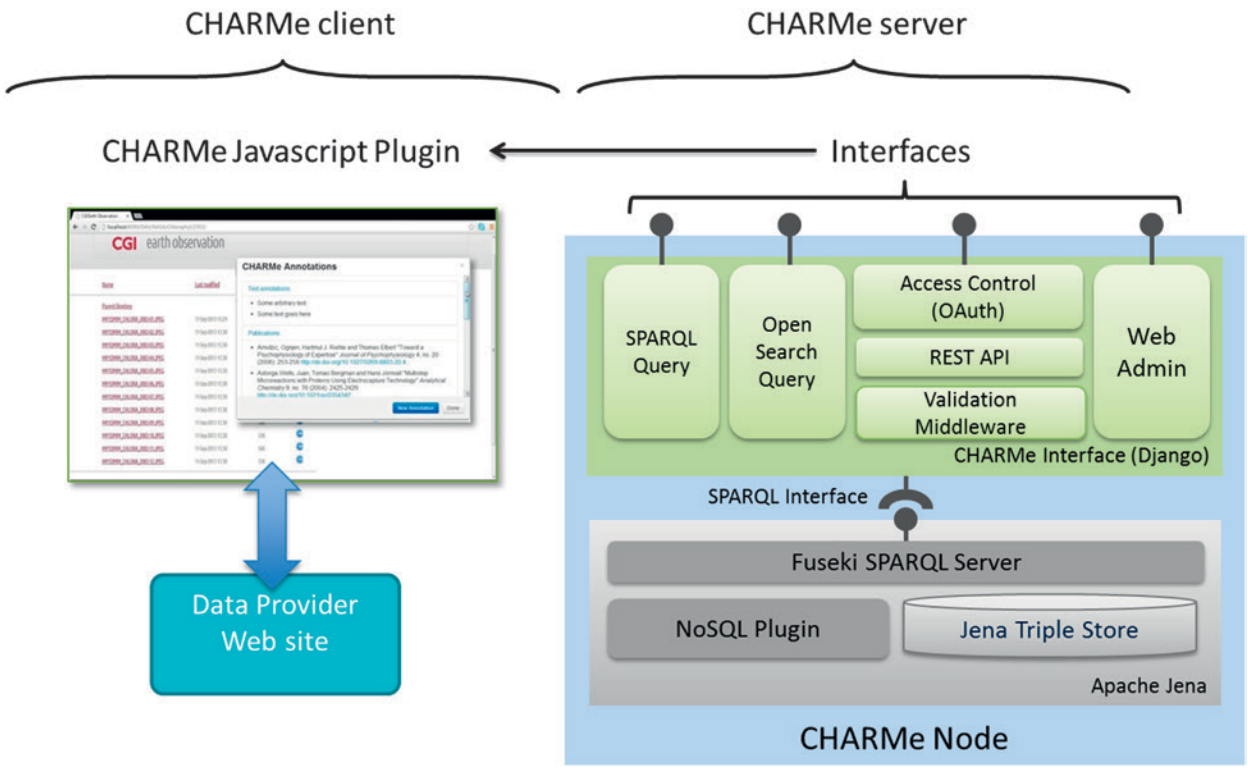


Fig. 4. Overview of CHARMe client-server architecture, with the JavaScript plug-in shown as an example of a client program. More details on the technologies used can be found in the technical documentation listed in the For Further Reading section. The CHARMe node (blue box) provides a range of web interfaces (green box). SPARQL and REST-ful (OpenSearch) query interfaces are provided for structured queries using RDF metadata. A REST API allows for submission, deletion, or modification of annotations and supports a JSON-LD serialization. The security layer provides access control via an OAuth 2.0 interface. Validation middleware checks the format of what has been submitted. A web admin interface allows users with the appropriate privileges to log in to the node directly (e.g., as a moderator) or to set up new data providers. The triple store and interface (gray box) is based on Apache Jena and Fuseki, respectively. The triple store is augmented with a NoSQL plug-in to index information and improve performance for search.

shown in Fig. 1, CHARMe also provides a convenient way to share information from the data provider (e.g., dataset provenance, updates, or corrections) at the point of access.

CHARMe has been implemented as a client-server architecture. On the server side, there is currently a single repository (a CHARMe "node") that stores all the annotation information and has interfaces to support many clients. Figure 4 shows the architecture of a CHARMe node, with the plug-in as an example of a client application. In most cases, the node does not store the target information itself—for instance, the actual dataset and conference paper journal in Fig. 3—but instead stores a link to the target; it is an important principle of Linked Data that these targets have persistent identifiers, such as a Digital Object Identifier (DOI) or persistent Uniform Resource Locator (pURL). This is critical not only for the CHARMe system to work but also for the larger problem of wider use of data citation in general.

The philosophy underlying the technical implementation was to develop a generic piece of software that can be configured to work with different underlying "off-the-shelf" technologies, and for the information held in the CHARMe node to be accessible via a range of web service interfaces. For example, CHARMe provides an open, read-only, web service endpoint (using the SPARQL protocol: www.w3.org/TR/rdf-sparql-query/) allowing potentially complex data mining and analysis for data scientists, as well as a simpler (but more limited) OpenSearch interface (www.opensearch.org/). The faceted search facility

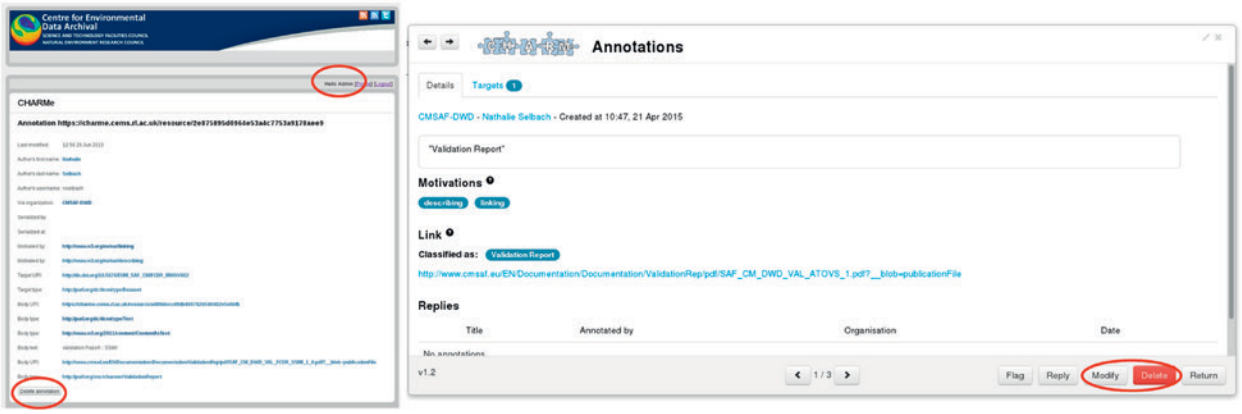


Fig. 5. The screenshots show a user viewing an annotation via (left) a browser or (right) the plug-in. If the user is the creator of the annotation, or has moderator or superuser privileges, the interface will include a "Delete" button, as highlighted, with the further ability to "Modify" an annotation via the plug-in.

in the plug-in is effectively a graphical user interface for queries to this OpenSearch interface. As these interfaces are built on open standards, it is possible for third parties to build other applications that produce and consume CHARMe commentary, with the CHARMe information syndicated to multiple end-user applications.

CHARMe uses the popular OAuth 2.0 framework (http://oauth.net/2/) to secure interactions between client programs and the node. The user authenticates and delegates permission to the client program to execute any secured operations on their behalf. All annotations submitted are publicly accessible in read-only mode; add, delete, and modify functions are secured and require login. Users have authority to modify or delete only the annotations that they themselves have submitted. In addition, however, there are two other elevated sets of privileges, for "moderators" and "superusers," which require registration with the node. Moderators are able to modify and delete any annotation originating from their client program(s), while superuser privileges can be granted to an overall administrator or administrators for the node. The client program is identified by an ID assigned by the node to the instance of the program (such as the plug-in) when it is deployed, and the moderator has oversight of content entered from their deployment of CHARMe. Superusers, in contrast, can modify or delete any annotation, from any source. This is provided as a second line of support for the moderation function used by individual clients. Example interfaces from a browser or the plug-in are pictured in Fig. 5.

EXPLORING COMMENTARY IN SPACE AND TIME. As part of the CHARMe project, the European Centre for Medium-Range Weather Forecasts (ECMWF) has developed a web-based graphical tool for associating features in climate time-series data with commentary that indicates "significant events" that may have affected the data, such as volcanic eruptions or satellite instrument failures. This "Significant Events Viewer" has been developed to work with ECMWF's reanalysis datasets and internal observation and events databases, but is designed to be general enough to be extended to other datasets and user needs. The viewer provides users with an opportunity to become more familiar with the variety of observations that feed into the reanalysis, and to determine whether the variability and features seen in the dataset are likely to be artifacts of the measurements or processing steps, or real changes in the environment. A registered CHARMe user can also add commentary to the significant event. Figure 6 shows an example of the Significant Events Viewer being used to explore ECMWF's reanalysis datasets ERA-40 and ERA-Interim alongside the significant events timeline. The tool is publicly available, following a simple and free registration, at http://apps.ecmwf.int/significant-events/.

To explore commentary in geographic space, CHARMe has developed a further prototype tool, "CHARMe Maps," that demonstrates the use of commentary metadata in an interactive mapping interface. The tool has two main capabilities, which are illustrated in Figs. 7 and 8:

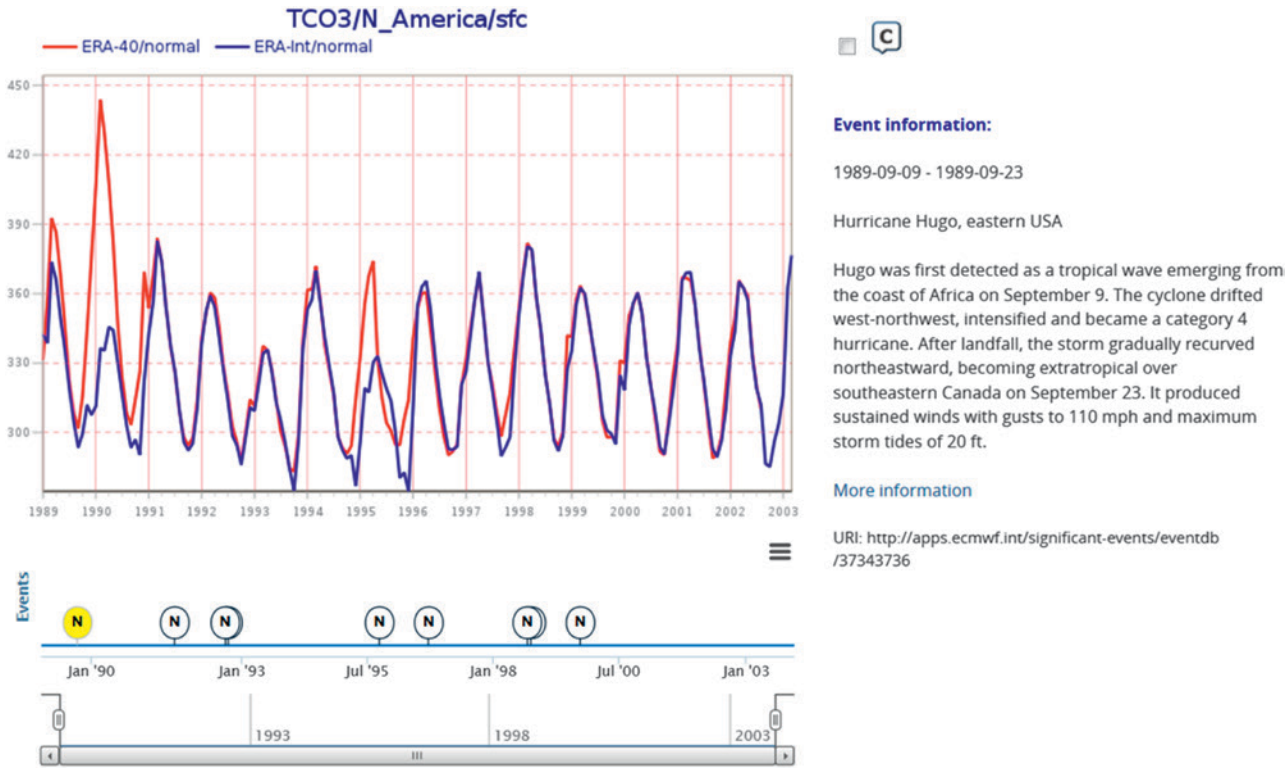


Fig. 6. The Significant Events Viewer being used to explore time series of global ozone in two reanalysis datasets (ERA-40 and ERA-Int), with a significant event timeline plotted below. The right-hand panel, "Event Information," describes the selected significant event (indicated by the yellow bubble). The clear CHARMe icon (top right) indicates that there are currently no user annotations recorded for the selected significant event.

1) the ability to attach annotations to a specific subset of a dataset—for example, a particular geographic region (we refer to this as "fine-grained commentary"), and
2) the ability to compare two datasets both visually and by the commentary that has been attached to them (we refer to this as intercomparison).

The linking of annotations to a subset of a dataset requires a modification to the general data model in Fig. 3. Fortunately, this kind of capability was anticipated by the designers of the Open Annotation specification. The properties of the subset include a geographic extent (defined as a 2D geometry), a temporal extent (defined by start and end times), and a vertical extent (not used in this prototype). Additionally, the definition of a subset allows the user to specify exactly which variables within a dataset are considered to be part of the subset; in this way, the user can attach a comment to a very specific part of a multivariate, multidimensional climate dataset. Figure 9 illustrates this data model for fine-grained commentary.

FUTURE PLANS FOR CHARME. The CHARMe project has built a system to support the creation, modification, and archiving of comments linked to climate datasets and other targets. The success of the CHARMe tools in achieving the vision set out at the start of this article will depend not only on further technology development, but also on the cultivation of a community of users who will build up the web of annotations and links over time. Although the project focuses on climate science, the technologies and concepts are very general and could be applied to other fields.

Data providers can enable CHARMe functionality in their websites by installing the JavaScript plug-in. It is also possible for institutions to host their own CHARMe node to store annotation information, but there is as yet no capability to federate searches across multiple nodes, so this is more appropriate if an institution wishes to keep annotations internal rather than public.

The node provides a standard API from which it is hoped many different client applications could be

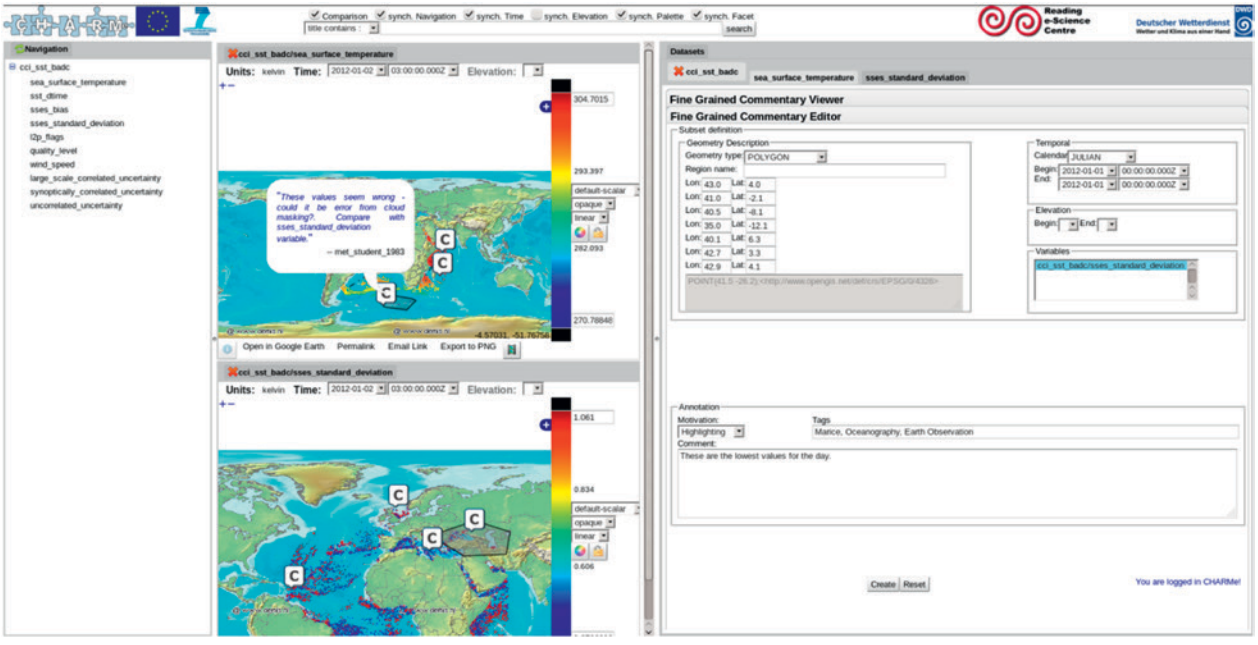


Fig. 7. A screenshot of the CHARMe Maps tool, highlighting the "fine-grained commentary" capability. Here, the user is visualizing two variables from a dataset (sea surface temperature and its associated error field) along with comments that have been attached to specific points or regions within the dataset. Note that each variable is associated with a different set of commentary.

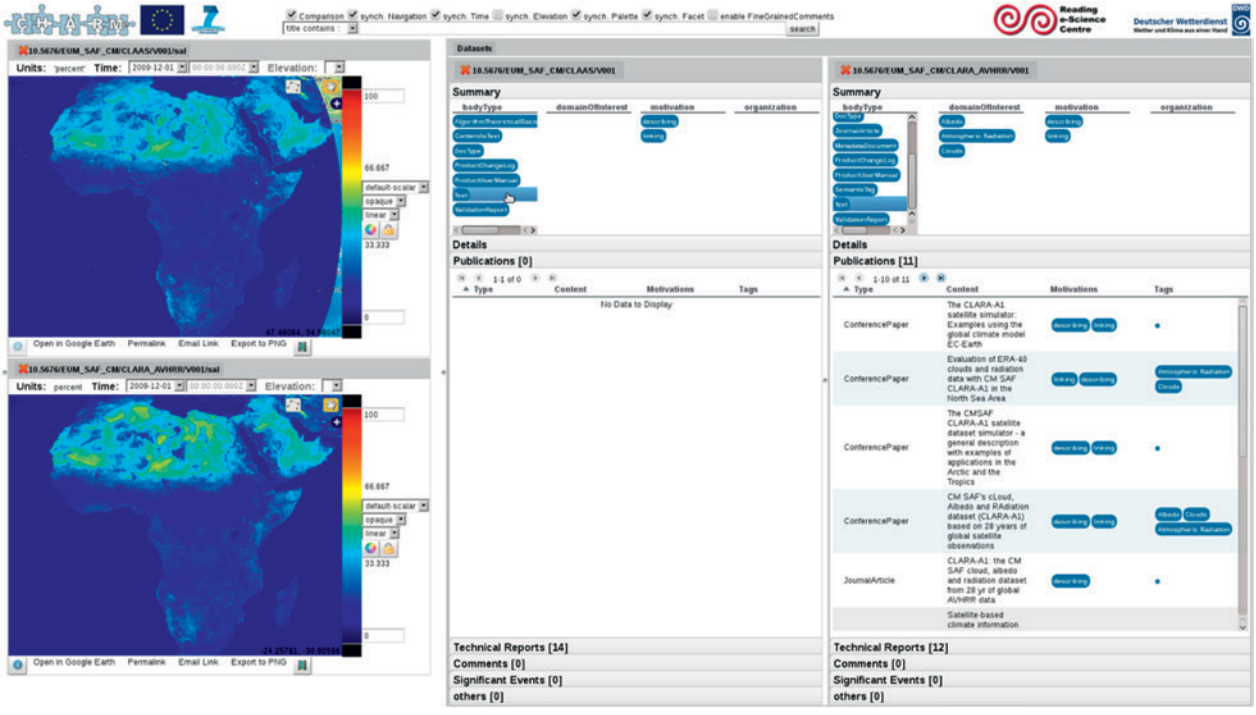


Fig. 8. A screenshot of the CHARMe Maps tool, highlighting the "intercomparison" capability. Two different albedo datasets are being visualized (left), with the two right-hand columns showing the commentary metadata that have been attached to each dataset. The dataset described in the right-most column (corresponding to the lower one in the visualization panel) has a number of publications attached to it.

developed, of which the Maps tool is just one example that demonstrates the possibilities. One potential reuse of CHARMe that is under consideration is integration into NASA's Giovanni tool, a web-based tool designed to enable visual data exploration and comparison of data offered by the Earth Observing System Data and Information System (http://giovanni.gsfc.nasa.gov/giovanni/). In the most recent Giovanni architecture, service workflows are specified as URLs that encode the service request and a specification of the data subset to be visualized, comprising the data variables, spatial region, and temporal range. That is, the URL includes the same information to specify a data subset as the specification in the CHARMe Maps data model. As such, the open machine-accessible architecture would allow a fairly straightforward incorporation of the CHARMe Maps ability to support commentary on data subsets.

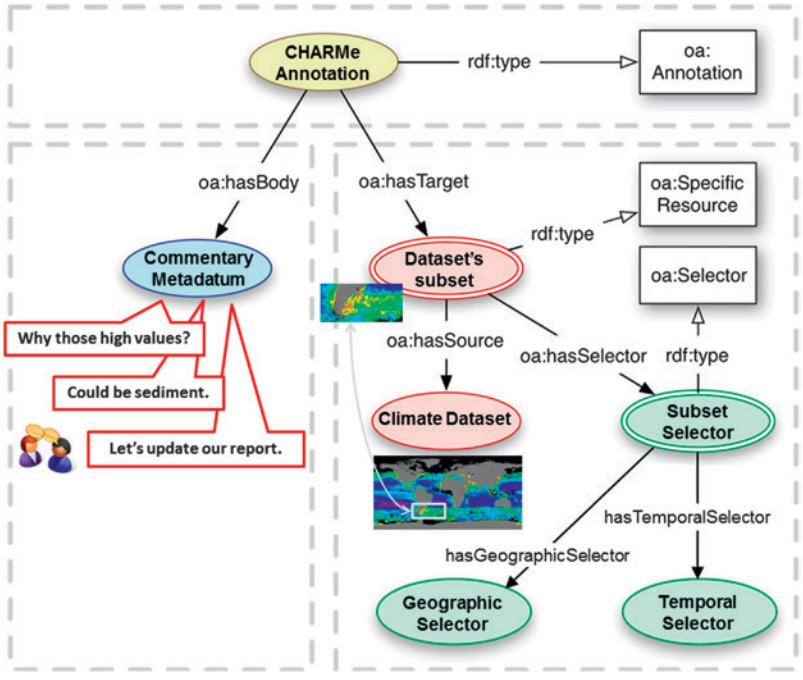


Fig. 9. Simplified representation of the data model for fine-grained commentary illustrating the use of Open Annotation's capability to annotate spatial and temporal subsets of a resource (in this case, a climate dataset). The full data model for fine-grained commentary includes more properties of the SubsetSelector.

Another potential NASA system to explore CHARMe integration is the related Regional Climate Model Evaluation System (RCMES; https://rcmes.jpl.nasa.gov/). RCMES is both a database for observations and an analytical toolkit allowing regridding, metrics computation, and visualization demonstrating comparisons of observations and climate model outputs. We envision an augmented RCMES allowing users to leverage CHARMe to explain the results of their model evaluation, and science workbench notes that are not currently captured in the RCMES tool. There is also strong interest in using CHARMe in the Earth System CoG collaboration environment (www.earthsystemcog.org/). CoG provides a search interface to the Earth System Grid Federation climate data archive, which houses climate model output and other widely used datasets, along with wikis, forums, and other tools for distributed discussion and analysis. Here, CHARMe could be integrated into the data search as a way to build an online knowledge base available at the point of download.

A further need for development is around moderation tools and policy. It does not appear that the social media world has solved the issue of controversial annotations. The main risk to a CHARMe annotation is not so much the controversy it sparks, as the possibility that the debate becomes irrelevant to the initial annotation and overwhelms substantive commenting, with the value of the commentary lost in the noise. This risk is likely relatively small in the beginning while the community is also small, but can become problematic as the community grows. We are currently surveying social media implementations and literature for promising approaches—such as upvoting/downvoting, reputation scoring, and sort/group mechanisms—to mitigate this risk.

The CHARMe code and user manuals are available at https://github.com/charme-project. The CHARMe system software is open-source, released under a BSD license, permitting future projects to reuse the source code as they wish. The Maps prototype is not currently accessible as an operational tool, but we would be happy to collaborate with anyone wishing to develop this capability in their own system.

ACKNOWLEDGMENTS. The CHARMe project was coordinated by the University of Reading, and project partners were Airbus Defence and Space, CGI, Deutscher

Wetterdienst (DWD), the European Centre for Medium-Range Weather Forecasts, the Royal Netherlands Meteorological Institute (KNMI), the Science and Technology Facilities Council, Terraspatium, and the UK Met Office. The project received funding from the European Union's Seventh Framework Programme for research, technological development, and demonstration under Grant Agreement 312541.

FOR FURTHER READING

Berrick, S. W., G. Leptoukh, J. D. Farley, and H. Rui, 2009: Giovanni: A web service workflow-based data visualization and analysis system. *IEEE Trans. Geosci. Remote Sens.*, **47,** 106–113, doi:10.1109/TGRS.2008.2003183.

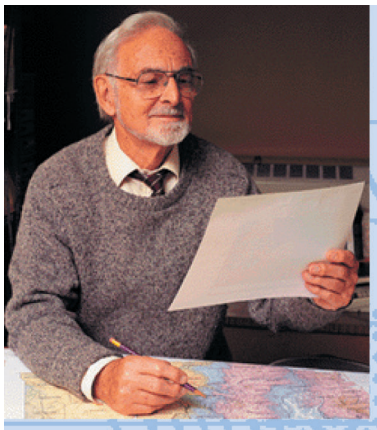
Blower, J. D., and Coauthors, 2014: Understanding climate data through commentary metadata: The CHARMe project. *Theory and Practice of Digital Libraries-TPDL 2013 Selected Workshops,* L. Bolikowski et al., Eds., Springer, 28–39, doi:10.1007/978-3-319-08425-1_4.

Dowell, M., and Coauthors, 2013: Strategy towards an architecture for climate monitoring from space, 39 pp. [Available online at www.wmo.int/pages/prog/sat/documents/ARCH_strategy-climate-architecture-space.pdf.]

Kershaw, P., 2014: Data model for commentary metadata (d400.1). Tech. rep., The CHARMe Project.

Nagni, M., and P. Kershaw, 2014: Concrete encodings of commentary metadata (d400.3). Tech. rep., The CHARMe Project.

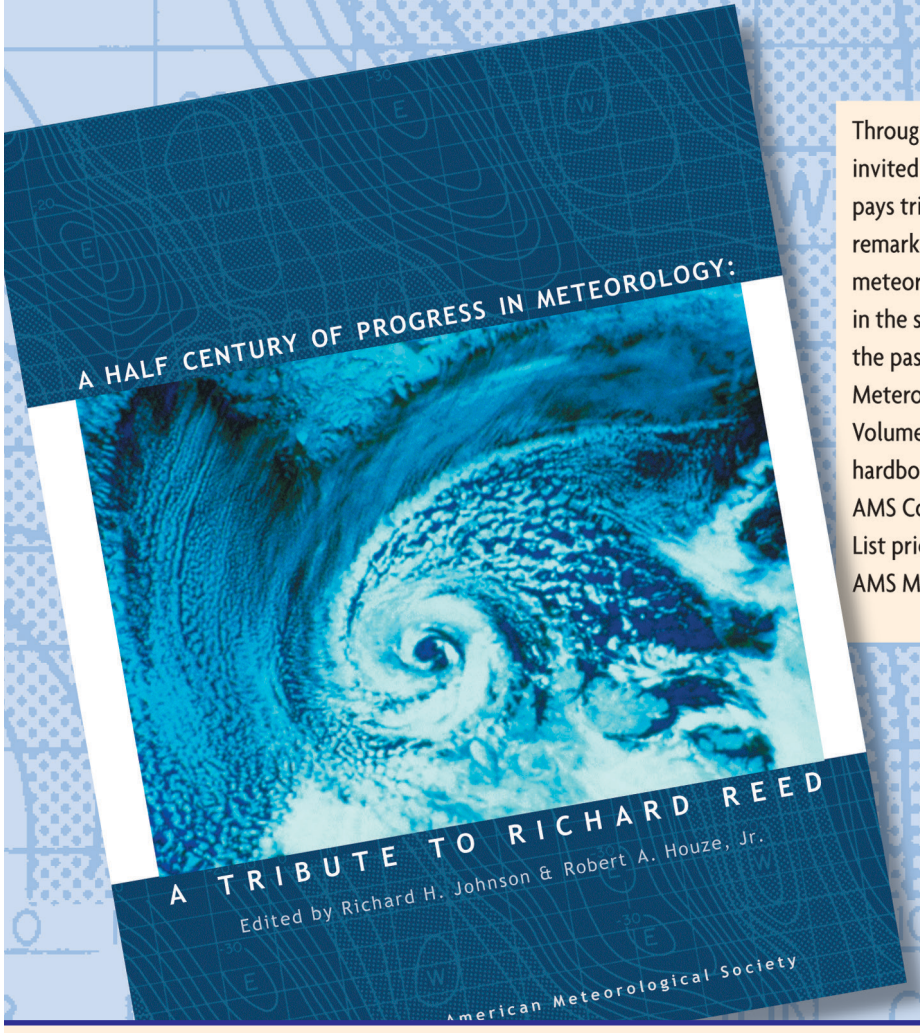
Rood, R., and P. Edwards, 2014: Climate informatics: Human experts and the end-to-end system. *Earthzine.* [Available online at www.earthzine.org/2014/05/22/climate-informatics-human-experts-and-the-end-to-end-system/.]



A Half Century of Progress in Meteorology: A Tribute to Richard Reed

edited by Richard H. Johnson and Robert A. Houze Jr.

with selections by: Lance F. Bosart Robert W. Burpee Anthony Hollingsworth James R. Holton Brian J. Hoskins Richard S. Lindzen John S. Perry Erik A. Rasmussen Adrian Simmons Pedro Viterbo



Through a series of reviews by invited experts, this monograph pays tribute to Richard Reed's remarkable contributions to meteorology and his leadership in the science community over the past 50 years. 2003. Meterological Monograph Series, Volume 31, Number 53; 139 pages, hardbound; ISBN 1-878220-58-6; AMS Code MM53.
List price: $80.00
AMS Member price: $60.00

ORDER ONLINE: bookstore.ametsoc.org or see the order form at the back of this issue