CLADDIER Report 1
User Experience of the prototype CLADDIER Discovery System

Katherine A. Bouton
National Centre for Atmospheric Science – Climate
University of Reading
Department of Meteorology
June 18, 2008

This report describes the user experience while using the prototype CLADDIER Discovery System. The users tested were climate scientists from the National Centre for Atmospheric Science - Climate, at the University of Reading.

Introduction

As part of CLADDIER, active climate scientist from the National Centre for Atmospheric Science – Climate, at the University of Reading were asked to provide feedback on the utility and ease of use of the CLADDIER Discovery System portal based on a number of documents placed in the repository. This feedback would provide CLADDIER partners with information on ways active scientists might actually use the system, how it might be improved, and what might need to be re-evaluated.

Testing Strategy

The users selected to participate in this report, were selected based on a determination by the Work Package Leader of their interest in a system like CLADDIER, and in their willingness to follow an interview strategy of questions which they would be asked to respond to.

The testing strategy included populating the CLADDIER database with data and publications based on the users initially selected to test the system as it was felt that this would make the prototype discovery system more interesting and understandable to the user. Note: Because this strategy of “pre-populating” the database with a few specific references, the user was provided with the exact search terms with which to test the system, preventing undue frustration during the testing phase.

A questionnaire was then written which was designed to take users through a series on steps to evaluate the CLADDIER project system. The selected users were asked to respond to the questionnaire, to evaluate and comment on the prototype discovery system. The responses have been compiled into this report.

Users were given the following instructions:

1. Go to the CLADDIER www site http://claddier.badc.ac.uk/trac to read more information on what CLADDIER is.

2. Go to the CLADDIER discovery portal by clicking on "online data and publication discovery service" (http://ignis.neodc.rl.ac.uk/claddier/search/single/).

3. Enter "search term" (the case of the search term doesn't matter) as your search term. You should get a red square and the title of your paper. This indicates that the search has come up with a single document referring to the search term.

4. Click on "Links: http://epubs2.cclrc.ac.uk/work-details?w=XXXX" which will take you to further details about the publications. At the bottom of the page is listed "Journal Articles". Click "view" under citations column in that table.

5. You should now have a window open which lists the Data as a "cited" dataset. That is a dataset that you had indicated you used in the paper. Note the format of the dataset citation.

6. Clicking on the link for that dataset http://ndgbeta.badc.rl.ac.uk/view/badc.nerc.ac.uk/Data should take you to the BADC data page for that dataset. On that page you will see a link to "cited-by: authors listed" and a link back to the paper.

7. The CLADDIER discovery round trip complete.

8. It should also work in reverse. Go to search start page http://ignis.neodc.rl.ac.uk/claddier/search/single/ and enter Data and you should be able to go through instructions above in reverse, ending back at the paper.

Types of questions you might answer:
1. Is this the kind of search work flow you would use
2. What kind of problems did you encounter
3. Would you use this type of data citation (in #5 above)
4. What problems do you foresee using this format
5. What you trust a reference of this sort
6. Any other comments welcome.

User Response

Users generally found that the instructions given in the questionnaire were successful.

The search workflow of the discovery system seemed to work fine for the users as they worked their way through the search process. Comments on the search workflow process included suggestions from multiple users that more work go into the search algorithms to recognize datasets with “obvious alternative dataset names” such that a search for ‘ERA40’ return the same results as a search for ‘ERA-40’. One user also commented that the search flow worked but only “to find other papers using the same data” and that this was of “only of real practical value if links to the papers online are included (eg doi)”.

All users agreed that there was a great deal of confusion over the terminology, particularly references versus citation. Users found that it was not clear what the 'view citations' button would do. All users wanted a clear distinction to find references to data or to papers. Would it show all the papers cited by this paper? Would it show the papers which cite this paper? More than one user suggested the complete phrase would be useful for navigation and that it be labeled "view dataset citations”. This

was especially true in the first stage of the search, causing confusion as to whether citations was citing all datasets used by that particular paper, or citing all papers which have cited datasets cited by that paper. Users did agree that it became clearer as the workflow progressed and a screen further into the search listed other papers who have also (supposedly) cited this (these) dataset(s). One user also suggested that this might cause problems when multiple datasets were involved.

One user mentioned using the help function for searching and that it did not include the explanation of any terminology.

Suggestions included:
A paper of uses DatasetA, DatasetB, and DatasetC.. A complete listing should produce 3 clear headings, one for each dataset, under which are listed published works that have cited those datasets.

Many users found that linking of datasets to papers, and papers to datasets was not correct, missing or "misleading". This was a critical issue for the users tested, and will be crucial in making such a system useable and useful to in the future.

Users were generally unclear about the data citation format. This may be because the examples used by the questionnaire from in the CLADDIER system were not adequate to test. All agreed the citation format should be as close to a journal citation as possible. One user wanted "some number (perhaps equivalent to doi) that would be appended to the paper details in the references section, and would be clickable. For example, if Smith and Jones (2008) released a new precipitation dataset, I could cite S&J (2008) as normal in the text. Then I could look it up in the references section, and find S&J (2008), Journal of Rain Gauges, p1-20, doi= [dx.doi.org link], dataset= [claddier_doi link]". Users did suggest they would use a data citation if it became standard to include data citations in the references of papers. However, as users noted, it is usual to cite a journal paper documenting the data. The problem with that is type of paper may not exist.

It was felt that an issue of primary importance for the dataset citation would be trust and the possibility of "assuming overconfidence in a given dataset because it is linked on claddier". Users indicated that trusting a citation would depend on the referencing – i.e. there are many slightly different versions of a particular Glocal Climate Model (GCM). Many problems would result from an incorrect reference given which would direct a user to a wrong version of the data. Users suggested that they may only trust those references from sites they were already aware of.

Conclusions and Recommendations

Generally users felt that search tools of this type would be very useful, "particularly for studies using a combination of several observed and model generated datasets, some or all of which were accessible through BADC or similar resource centres (e.g. ESRL at NOAA)." The portal itself requires more work to make it useful. The correct links from data to papers must be made. The search function must be improved and the terminology used in the results must be made clearer.

Great effort and care would need to be expended such that the citations, references, and links to datasets were correct and trustworthy. How that process can be automated, and how that "trustworthiness" was to be gained remains an issue.

Summary Recommendations

1. Useful alternative search terms would help (e.g. variants on era40).
2. All references and citations should be hyperlinked.
3. There needs to be clarity in the interface as to the distinction between references and citations.
4. Alternative presentations of the citation graph would be desirable (to clearly show more than one step in the process: e.g. from PaperA, show references to DatasetB and DatasetC and show the papers which cite DatasetB and C immediately. In this case keywords would help disambiguate between the reasons why specific papers were cited.
5. The system must be reliable and trustworthy. Experience of incomplete and incorrect results would result in little usage.
6. The overall usefulness would depend on reliable data versioning and documentation to ensure that citations point to trustworthy data sources. (This is consistent with the CLADDIER view on data publication)