ATARRABI – A WORKFLOW SYSTEM FOR THE PUBLICATION OF ENVIRONMENTAL DATA

Florian Quadt[1], André Düsterhus[2], Heinke Höck[3], Michael Lautenschlager[3], Andreas V. Hense[1]*, Andreas N. Hense[2], Martin Dames[1]

*[*1] Dept of Computer Science, Bonn-Rhein-Sieg University oAS, Grantham Allee 20, 53757 Sankt Augustin, Germany*
Email: {andreas.hense|florian.quadt|martin.dames}@h-brs.de
[2] Meteorological Institute, University of Bonn, Auf dem Hügel 20, 53121 Bonn, Germany
Email: {ahense|andue}@uni-bonn.de
[3] German Climate Computing Center, Bundesstraße 45a, 20146 Hamburg, Germany
Email: {lautenschlager|hoeck}@dkrz.de

ABSTRACT

In a research project funded by the German Research Foundation, meteorologists, data publication experts, and computer scientists optimised the publication process of meteorological data and developed software that supports metadata review. The project group placed particular emphasis on scientific and technical quality assurance of primary data and metadata. At the end, the software automatically registers a Digital Object Identifier at DataCite. The software has been successfully integrated into the infrastructure of the World Data Center for Climate, but a key was to make the results applicable to data publication processes in other sciences as well.

Keywords: Data Publication, Environmental Data, Meteorological Data, DataCite, Workflow Management, e-Research, Digital Object Identifier, DOI

1 MOTIVATION

In these days, research is conducted in a highly computer-aided manner, and it has become more and more challenging to select, annotate, store, and publish the data generated in order to enable other people to find and reuse them for further research (Science Staff, 2011). Besides the technical aspects, scientific quality is a very important factor for determining the scientific value of the data. An example of this kind of research is the field of meteorology, where vast amounts of data are produced in observations and simulations. These data have to be stored and published in a way such that they can be uniquely cited in publications and easily retrieved.

Technically speaking, persistent identifiers are already in place, e.g., those of the Handle system or Digital Object Identifiers (DOIs). These identifiers consist of a unique, immutable resource name and a landing page that contains or references the actual content. Both are registered in a central and publicly available directory. If the data move to another site, the landing page value can easily be changed to the new location. Persistent identifiers have been used for scientific articles for some time now and have become increasingly popular for citing research data as well (Paskin, 2005 ; Brase, 2004). Just to mention an example, the citation reference to data which have been assigned the DOI "10.1594/WDCC/dphase_mpeps" looks as follows:

> Denhard, Michael (2009): dphase_mpeps: MicroPEPS LAF-Ensemble run by DWD for the MAP D-PHASE project. World Data Center for Climate. http://dx.doi.org/10.1594/WDCC/dphase_mpeps

The Internet address has been set up on the basis of the DOI, and it enables the reader to directly interpret the DOI and to be directed to the current landing page.

Since the availability of research data is an important factor for reuse, the data usually have to enter a long-term archive as a prerequisite for being attributed a persistent identifier. In addition to this logistical requirement, the data themselves have to be prepared for being cited and used by other researchers and thus have to be of high scientific

and technical quality (Hense & Quadt, 2011). The process to ensure these quality aspects differs from one institute to the next and is hardly systematised and automated.

From April 2009 to March 2012, three complementary institutions conducted the joint research project "Publication of Environmental Data" (http://umwelt.wikidora.com). The project was funded by the German Research Foundation (DFG), with the aim of developing a well-defined scientific data publication process (based on research data that had already been stored in a long-term archive); the project was also aimed at developing a software solution that supports and automates this process. The following roles were represented:

- meteorologists as subject matter experts for the publication of data from observational experiments,
- experts for the management of (meta)data and the data publication and DOI registration procedure, and
- business process analysts and software developers.

In the following, the approaches to and achievements of the research project will be presented in detail. In Section 2, the data publication process is outlined in general terms and described in the context of the field of meteorology; data publication is compared to traditional text publication at this point. Section 3 deals with data publication from the three complementary perspectives of the individual partners, namely those of the scientist, the data centre, and the process automation staff. The software solution finally developed is introduced in Section 4. Section 5 gives an outlook on how the ideas and concepts can be further developed, and Section 6 concludes with the summary of the findings.

2 INTRODUCTION TO SCIENTIFIC DATA PUBLICATION

2.1 Data Publication in General

This section gives an overview of the procedure used to publish research data. The general data life cycle includes the following steps: production, evaluation, archiving, and dissemination (Lautenschlager, 2011, cf. Figure 1). The first three steps cover the procedures from measuring or generating data up to entering the data into the long-term archive.

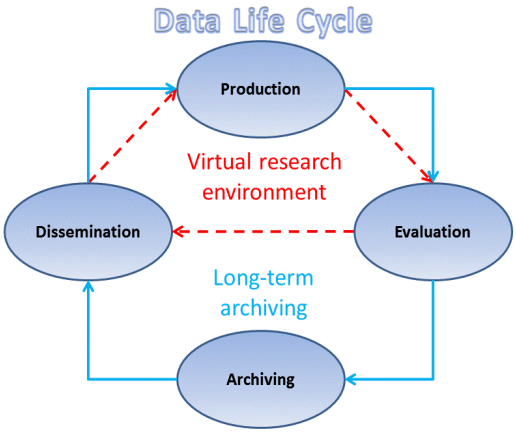

Figure 1. The data life cycle as defined by Lautenschlager (2011)

Dissemination involves further activities needed to expose data, such as registering a persistent identifier for the data and making the data visible to search engines. In the course of the project, the process of selecting and preparing data for publication and finally assigning a DOI was analysed, optimised, and systematically documented as a business process model. Three generic roles or parties that participate in the publication process were identified:

1. The **scientist** is an author or another authorised person familiar with the primary data and metadata. He or she reviews the primary data and metadata required for publication (scientific quality assurance, SQA) and provides additional documentation (e.g., plots, notes about the quality assurance measures performed).

2. Each data publication procedure is also assigned to a **publication agent** who is a representative of the publication agency. He or she supervises the process, supports the scientist in the event that any problems should arise, and double-checks the data entered. Additionally, the publication agent performs checks on technical criteria, such as completeness of the datasets (technical quality assurance, TQA).

3. The **registration agency** is an organisation that is able to register new persistent identifiers and to update those already in place, such as DOIs or URNs. These agencies normally provide an interface that can be used for machine-to-machine communication, e.g., a web service to register a new identifier.

In order to structure the publication process, it was divided into the following four sub-processes (cf. Figure 2):

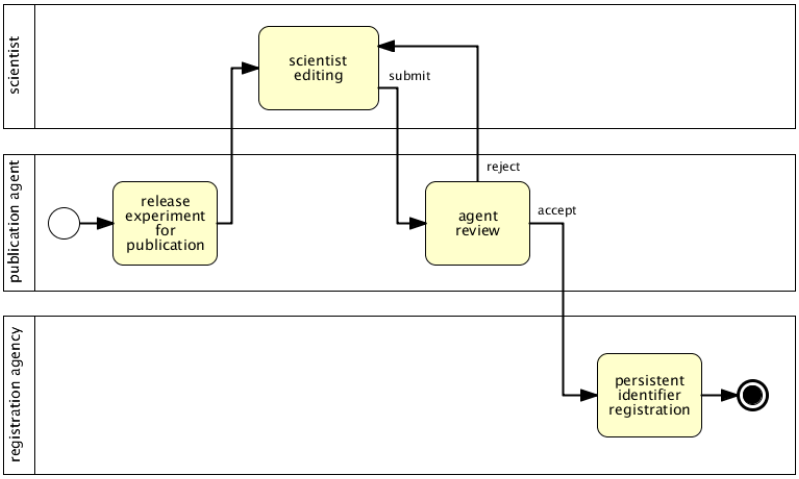

Figure 2. The general process of data publication (simplified)

1. **Releasing experiments for publication**
   Not all archived experiments are suitable for data publication. Before the publication process of an experiment can start, a publication agent has to explicitly release that experiment for publication. At the same time, a scientist who will conduct the publication process is designated.

2. **Scientist editing**
   The scientist performs the scientific quality assurance checks. In the course of this phase, he or she ensures high quality of the primary data and reviews and edits all the metadata required for publication. The long-term archive provider may have specified additional metadata that have to be reviewed or added. After the scientist has completed the metadata review, he or she releases the data for review by the publication agent.

3. **Agent review**
   After the scientist has submitted the metadata, a publication agent performs the agent review by double-checking all the metadata and performing the technical quality assurance measures. At the end of the review, the agent has to decide whether he or she rejects or accepts the data submitted. In the former case, the scientist will have to revise the metadata and re-submit them. In the latter case, the publication process proceeds to the identifier registration step.

4. **Persistent identifier registration**
   To register a persistent identifier and the mandatory metadata (landing page, title, authors etc.), an XML file complying with the registration agency's interface is generated and passed on to the registration service.

Thus, the process ends with the public assignment of a persistent identifier and the specification of some essential metadata (e.g., authors, title, year). These elements together form a citation reference, which can then be used to cite primary data in paper publication. Scientific data published this way is marked as irrevocable. As a consequence, error corrections result in a new version of the data along with a new data publication instance.

2.2 Data Publication in the Field of Meteorology

Based on the general explanations in Subsection 2.1, the process will now be defined in more detail for the field of meteorology, and the developed software system, termed "Atarrabi" (Basque: good weather spirit), will be embedded in the context of the publication process.

In the case of meteorological data, the publication agency is the World Data Center for Climate (WDC Climate) in Hamburg, which is staffed by the German Climate Computing Centre (DKRZ), operating high performance computers and the long-term archive CERA for this kind of data. The WDC Climate offers a publication service to register a DOI and a URN for data that have already been archived in CERA. The identifiers are registered via Technische Informationsbibliothek (TIB), Hannover. Since TIB is a member of the DataCite (http://www.data-cite.org) consortium, the technical interfaces of DataCite are used to register new DOIs and URNs.
Figure 3 illustrates the way the various parties collaborate and Atarrabi is integrated into the organisational structure. One can easily recognise the scientist performing the SQA task, the publication agent at WDC Climate performing the TQA task, and the registration process at DataCite. Atarrabi supports and integrates these process steps and oversees some additional activities:

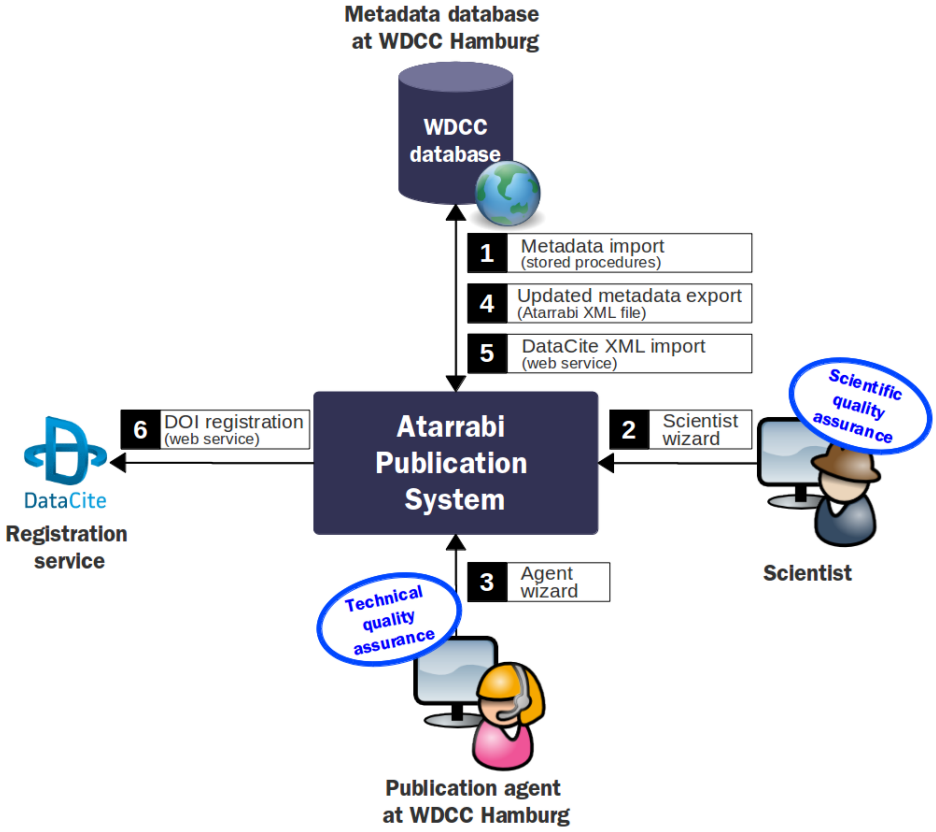

Figure 3. Workflow of the Atarrabi Publication System

1. Atarrabi imports all relevant metadata of the meteorological experiment to be published from the CERA database at WDC Climate Hamburg.

2. The scientist reviews and edits all the metadata required for DOI/URN registration. The software component used to carry out this process step has been named 'scientist wizard' and displays all elements

of the DataCite kernel, such as the related authors or contributing institutions, the title, date, etc. Atarrabi validates the data entered and assures that all mandatory fields are filled in. If the scientist has a comment or a question regarding the metadata, he or she can use a dedicated text field to communicate with the publication agent. Since it was a main objective of the project to emphasise quality assurance and its documentation respectively, fields for this kind of information have been added to the data entry forms in the wizard as well. After the scientist has completed the metadata review, he or she finishes the scientist wizard by submitting the data.

3. After the scientist has submitted the experiment data, the publication agents are informed that a new experiment is awaiting review by an agent. One of the publication agents can now assign this task to him- or herself and start the review ('agent wizard'). Similarly to the scientist wizard, the agent double-checks all the metadata submitted by the scientist. Additionally, the publication agent is able to enter information about technical quality checks, such as the check sums and other technical criteria (TQA). Having reached the end of the wizard, the agent has to decide whether he rejects or accepts the metadata and research data provided by the scientist. In the former case, the scientist will have to re-work the data accordingly and re-submit it.

4. If all checks have been passed successfully, the metadata reviewed are exported back to the CERA database, which is updated accordingly.

5. An XML file is now generated by CERA, which contains the mandatory metadata (DOI string, landing page, title, authors, etc.) and is compliant with the DataCite XML schema for DOI registration. This file is imported into Atarrabi.

6. Atarrabi calls the DataCite registration web service using the XML file imported previously. This finally registers the DOI and the mandatory metadata in the DataCite database. If this call is successful, the publication process is complete.

To learn more about the appearance of the system and the user interface concepts of Atarrabi, please refer to Sections 3.3 and 4.

2.3 Data Publication Compared to Text Publication

The first difference to mention when comparing data publication to traditional text publication is related to the contents and the formats of the data files used and the impact of these file types on storage and data search. Table 1 gives an overview of some of the differences, as presented by Hense & Quadt (2011).

Table 1. Selected differences of the file types used for storing text and data publications

	Text	Data
File formats	few widely accepted file formats	many, discipline- and tool-specific
Contents	optimised for human reader	sometimes difficult to inspect
File size	mostly moderate (less than 2 MB)	can be huge
Scientific quality assurance	peer-review long tradition	no peer-review (yet?)
Formal and technical quality assurance	orthography, typesetting, etc.	partial automation possible
Browse & search	metadata plus full text search	often restricted to metadata
Storage site	single file, stored directly in repository	repository only; stores links to huge files

While this perspective on data and text publication is rather technical, differences within the publication process itself can be identified as well. Figure 4 shows the traditional publication process of a written document (see blue boxes). This process may start with the development of hypotheses in pre-experimental theory. These hypotheses lead to predictions, which should now be validated by performing an experiment. The scientist designs and executes the experiment and evaluates the data gathered, e.g., by carrying out statistical analysis in accordance with post-experimental theory. This theory leads to an analysis and interpretation of the data in the context of the whole experiment. Based on this, the scientist can write a manuscript and submit it to a journal, where it will be peer-reviewed to ensure scientific quality. Provided that the outcome is positive, the work will then be published as a written document.

As a consequence of the new opportunities provided by information technology, this process has been improved over the years and been extended to deal with primary data in publication (see orange boxes in **Figure 4**Figure 4). At first, some publishers started to include a quality assurance procedure outside the traditional peer review process by introducing discussion papers (Pöschl, 2010) or institutionalising pre-publications (arXiv) (Gura, 2002). Data papers, such as ESSD, have also been developed (Pfeiffenberger & Carlson, 2011). These do not include the full scientific workflow summarised in Figure 4. Nevertheless, they explain the main parts of the experimental design as well as the experiment itself, and they also perform preliminary checks or evaluations with respect to the experimental design of the dataset. Since both are free-form text, the quality assurance measures and the peer review process are similar to those performed in traditional paper publication.

Data publication, finally, is the third method used to publish experimental results for independent and interdisciplinary third-party use (see green boxes in Figure 4). Data publications only include the data and an extensive set of metadata. The latter summarises the main information about the experiment and the structure of the data. Both parts need quality assurance, which can generally be performed by the author himself, a data centre, or under the guidance of a scientific journal. In the research project, an author-based approach is defined, which is depicted in Section 3.1.3. Additionally, the metadata have to be brought to an appropriate standard in the course of the quality control procedure. After this process has been completed, the data can be published and cited in the data paper or, even more importantly, in traditional text documents.

A peer review for data (see black box in Figure 4) would also be a logical consequence and an important future step. However, how to do this effectively for (large amounts of) primary data is actually a matter of ongoing research (Parsons, Duerr, & Minster, 2010; Lawrence, Jones, Matthews, Pepler, & Callaghan, 2011).

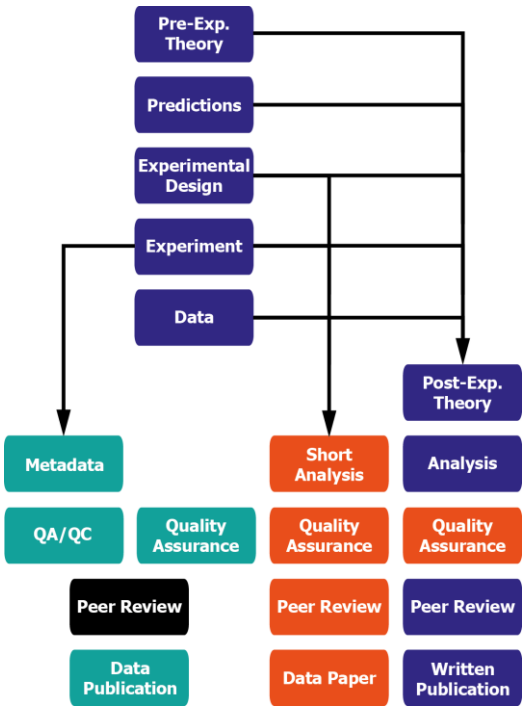

Figure 4. A schematic description of an experiment-driven scientific process, including the different ways of publication. It shows the traditional publication process (blue), the new processes made possible through the advent of the World Wide Web (orange), and the data publication process (green).

3 THREE VIEWS ON DATA PUBLICATION

3.1 The Scientist's Perspective

3.1.1 Data Classes and Formats in the Meteorology Domain

In atmospheric sciences or climate research, one can distinguish two major classes of research data: simulation data and observational data. Simulation data are the output from meteorological or climatological computer models and are thus characterised by a time and space structure in the computational grid (due to this regular structure, data from satellites can be included in this category as well). For a spatially extended and instationary model, there are obviously the three dimensions in space and the time dimension. Additionally, fluid dynamical codes, such as climate or weather models, provide various physical variables, such as velocities, temperatures, and pressure. Climate and weather models are typically nonlinear models showing the typical sensitivity to small uncertainties in the initial conditions (the well-known "butterfly effect" (Lorenz, 1963)). Monte Carlo simulations are the relevant tools used to explore these sensitivities. On the whole, this leads to a well-defined six-dimensional data model, including the three space dimensions, time, a variable, and the realisation dimension, respectively. Although it is possible to re-calculate a model if all components of the simulation (code, input data, operating system, compiler, etc.) are available, this might lead to an unacceptable overhead (Palmer, 2012). Therefore, it is generally reasonable to archive the output from complex simulations.

Numerical weather predictions have a fifty-year-old tradition, and global climate models have been compared systematically for about 20 years. Therefore, the data formats of these data have been highly standardised by the model community (Meehl, Covey, Taylor, Delworth, Stouffer, Latif, et al., 2007), e.g., Gridded Binary (GRIB) and netCDF. The GRIB (http://www.wmo.int/pages/prog/www/DPS/FM92-GRIB2-11-2003.pdf) representation format has been initiated and standardised by the World Meteorological Organisation (WMO). Within the framework of the World Climate Research Program, the Coupled Model Intercomparison Projects (see Subsection 3.1.4) have

propagated the netCDF standard. It is especially the latter standard that allows for the documentation of the metadata relating to the data.

The second class of data is observational data; they can be further divided into two subclasses. The first subclass comprises all permanent networks of different types. These produce observational data that are used as input for meteorological weather forecasting and that form the basis for climatological analysis on a longer-term scale. Some of these networks provide observational data on a regular basis in space and time, such as the official meteorological surface stations or radiosonde stations, operated mainly by national weather services. Other permanent networks, such as aircraft observations or polar orbiting satellites, collect data irregularly distributed in space and time. In recent years, remote sensing networks, such as Radar, have been installed, which produce area-filling and sometimes volume-filling data on a 24/7 basis. Almost all of these data are stored in standardised formats. Some of these formats have been used for more than forty years, such as the international Binary Universal Form for the Representation of Meteorological Data (BUFR) or, again, the GRIB format. The reason for this can be found in the necessity to include these data as an input parameter during the numerical computation and preparation of initial conditions for numerical weather prediction models.

The second subclass comprises observational datasets that are produced by non-permanent networks (e.g., installed during field campaigns). Very often datasets from field campaigns are self-contained due to their uniqueness, which calls for further attention, e.g., the initiation of new questions and research. The standardisation of methods for performing measurements and storing data are less common here, and the measured data itself can be of much higher complexity. This can be due to the fact that experimental and non-standard measurement techniques are used during the experiment. There is a wide variety of space-time variable structures, ranging from pointwise observations of time series of a large number of chemical constituents through to measurements of state variables in the atmosphere and in oceans, taken along trajectories by drifting observation platforms, such as aircraft, air ships, ships, or buoys. The measurements can be area- or volume-filling, e.g., by using active or passive remote sensing instruments at a fixed location in space or even on a moving platform, such as aircraft or ships. Due to their experimental status, the methods to be used for storage and data formats are often dictated by the modus operandi of the experimentalist and the special aims of the field campaign. Increasingly, all participants of a field project sign what is known as data protocols. These regulate the exchange and also the sharing of data among the participants and third parties, specify a data format such as netCDF, and set the related naming conventions for the climate and forecasts (CF) (Eaton, Gregory, Drach, Taylor & Hankin, 2011). Funding agencies, especially those which provide public finance, demand such protocols as part of their funding regulations.

3.1.2 Data Publication

If datasets from any of the data classes described above are available, their integration into a long-term data archive or data centre, as a logical consequence, enables data provision for subsequent research. In these days, this measure is even a frequent requirement imposed by funding agencies. The first archived datasets were collected in the course of the campaigns carried out during the International Geophysical Year in 1957/58, showing that meteorological and geophysical communities were at the forefront of this process. As a result, an increasing number of World Data Centers (WDC) for specific domains, coordinated by the International Council for Science (ICSU), has been established over the past years. This finally led to the creation of the World Data System (WDS) (http://www.icsu-wds.org). The participating data centres guarantee technical and organisational long-term availability of the archived data, e.g., through institutional funding and excellent expertise in hard- and software.

After the data have been stored in a common format and archived in a common database, the next objective is to improve the visibility of the data. The findings of a poll among about 1,700 scientists concerning their research practices were published in a recent issue of Science (2011). One interesting finding of this survey is that more than half of the scientists who participated rarely access data from published literature or data archives. Moreover, the survey reveals that about half of respondents exclusively store research data in their own lab and another 30% on university servers. These data are destined not to be found and used by other researchers. This insight calls for new strategies and standard procedures to improve data visibility, to simplify data citation and sharing, and to enable the author to earn credibility for his or her work (Costello, 2009; Henneken & Accomazzi, 2011).

Today, publishing in trustworthy publications is a crucial factor in earning credibility in the research community. This is especially important for researchers who are focused on measurements since it is very challenging to earn

credibility by publishing papers in traditional journals, which are mostly focused on subsequent data analysis. Even if the authors are able to publish their papers, they might want to underline the importance of their data by sharing them with other scientists on trusted platforms. For researchers who are focused on data analysis rather than data measurement, the importance of data publication can be found to lie in a simplified and transparent access to the data. They are able to obtain and use the data from official repositories by simply citing the original datasets. This is an important factor for future work in climate research, as stated in the review of the IPCC process (Committee to Review the Intergovernmental Panel on Climate Change, 2010) by the InterAcademy Council in 2010.

A promising approach to the simplification of data citation is the registration of a persistent identifier for the data, such as a Handle, DOI, or URN, which can then be quoted in citation references of scientific articles. The corresponding resolvers return the current data location for a given identifier based on a (mutable) location entry in the identifier catalogue. Using such a directory approach enables authors and data managers to move data among several data centres while retaining the integrity of the identifiers. Besides the cost for registration, the data publication process necessitates considerable effort, especially with respect to ensuring high quality of the primary data and metadata. Subsection 3.1.3 goes into the details of scientific quality assurance while Subsection 3.2.4 deals with the technical aspects.

3.1.3 Scientific Quality Assurance

In regard to Open Access initiatives, delivering quality-controlled data is essential. Both data availability and data quality are important in getting the best out of data measured by the scientific community. For scientists who want to use datasets provided by other researchers, this is especially important; drawing conclusions from the data is only possible if the quality of the datasets is known. This is not only restricted to a technical perspective of the files but also covers the content of the data. Therefore, it is generally possible and necessary to distinguish between Technical Quality Assurance (TQA) and Scientific Quality Assurance (SQA). The TQA procedure assures the integrity and completeness of the original datasets and their metadata after their transfer into the long-term data archive. This part of quality assurance is to a large extent part of the work performed within the data centre and the archiving process.
The SQA procedure can be divided into three parts. The first part includes the SQA check on the metadata (SQA on Metadata), which should include all relevant details relating to the production and processing of the original dataset. At this point, it is important to know who did the measurement, at what time and in what place the measurement was done, what the basic parameters of the measurement were, and what measurement devices were used.

The second part deals with data quality itself (SQA on Data). At this point, the major task involves looking for and evaluating indications of measurement errors, such as unphysical outliers or internal inconsistencies. In the third part, the measuring methods are described (SQA on Methods). These descriptions are more detailed than the standardised metadata attributes and may include instrument calibration, the precision of the time and location data, the situation around the measurement, or technical explanations of the measured parameters. All these details should be published as a printed document in a (specialised data) journal. In case of the SQA check on metadata and data, information acquisition and storage can be carried out directly at the data centre.

The simplest part of the SQA procedure with respect to standardisation is the SQA check on the metadata. Here, the main focus lies on authorship information and measurement details. The latter consist of information on location and time, the instruments used, the level of data processing, and a general description of the dataset. All this is important because a data user needs to know whether or not he or she can use the published dataset for research purposes.
This is also important for the SQA check on the on the data. In this case, it is reasonable to focus on a strictly test-based point of view. These tests should give insights into whether the dataset is reliable and which parts of the dataset need special attention by the potential data user. For this purpose, it is important to give a description of the tests used, some information on the algorithms used and their results, and most importantly, a comment by the author. The latter is necessary to help the data user in classifying the quality tests and in explaining their findings.
All this requires some effort on the part of the author, but it is very important because only quality-controlled data can be subsequently used by the data users. The extension package “qat” (Quality Assurance Toolkit) (http://cran.r-project.org/web/packages/qat) for the free statistical programming language R was developed for this purpose. It is able to perform quality tests on time series and to document their findings. In addition, new general quality tests have been developed to check the datasets (Düsterhus & Hense, 2012).

3.1.4 The Coupled Model Intercomparison Project (CMIP5)

The Coupled Model Intercomparison Project (CMIP) defines a standard experimental protocol for studying the output of coupled ocean-atmosphere general circulation models. It provides a community-based infrastructure in support of climate model diagnosis, validation, intercomparison, documentation, and data access. This framework enables a diverse community of scientists to analyse these circulation models in a systematic fashion, a process which serves to facilitate model improvement. The CMIP has passed through several phases. Currently, all CMIP activities are part of the fifth phase (CMIP5) (Guilyardi, et al., 2011), and the findings will be addressed in the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) scheduled for publication in 2013.
The model output data generated in CMIP5 simulations and the corresponding model and experiment documentation are or will be publicly available in the data portals of the Earth System Grid Federation (ESGF). These data portals are provided by several data centres, the World Data Center for Climate being one of them.

3.2 The Data Centre's Perspective

3.2.1 The World Data Center for Climate

The World Data Center for Climate (WDC Climate) (http://www.dkrz.de/daten-en/wdcc) is run by the German Climate Computing Centre (DKRZ) in Hamburg. The mission of WDC Climate is to serve the scientific community by collecting, storing, and disseminating research data with emphasis on climate modelling and related data products. There are plans to cooperate more closely with thematically related data centres such as Earth observation, meteorology, oceanography, paleoclimate, and environmental studies, with the aim of establishing a complete network for climate data.

Thanks to its integration into the DKRZ organisation, the WDC Climate has access to a professionally run computing and storage infrastructure (http://www.dkrz.de/Klimarechner-en). There are high-performance computers such as an IBM Power6-based cluster, which attains a peak performance of 158 Teraflops. To ensure a continuous flow of climate data between the high-performance computers and the archive system, DKRZ runs one of the largest (more than 100 Petabyte total capacity) and most powerful data archives in the world.

The WDC Climate has been a regular member of the ICSU World Data System (WDS) (http://www.icsu-wds.org) since 2003. The WDS is aimed at ensuring long-term stewardship and the provision of quality-assessed data and data services to the international science community and other stakeholders.

3.2.2 The Long-term Archiving Service

If a project has expired, data processing and evaluation have been completed, or the data are no longer needed for online access, the data qualify for the WDC Climate Long-term Archiving service (LTA) and subsequently enter the long-term data archive. Logically, the data are inserted in the high-capacity Oracle database "CERA" and thus become part of the WDC Climate. The voluminous primary data are saved as flat files or as container BLOBs (Binary Large OBjects). The metadata are stored in a specialised database schema to make them easily accessible to search engines (Lautenschlager, Toussaint, Thiemann, & Reinke, 1998). Physically, the data are stored on tape in a high performance storage system that maintains the online availability of the data (Toussaint, Lautenschlager, & Luthardt, 2007).

To browse and search the metadata and download primary data in the long-term archive, the WDC Climate provides a web-based user interface (http://cera-www.dkrz.de/WDCC/ui/Index.jsp). All downloads are free of charge, but the user needs to have a user account to authenticate him- or herself. For external groups and projects, the LTA service is available with payment of a fee.

3.2.3 The Data Publication Service

Once the data have entered the long-term archive of WDC Climate, they are indeed publicly accessible, but the datasets are still not known to the wider public. Moreover, the data lack a formal notation for being directly referenced in scientific articles. To fill this gap, WDC Climate provides a publication service for climatological data, which marks the datasets concerned as immutable, assigns them persistent identifiers (Duerr, Downs, Tilmes,

Barkstrom, Lenhardt, Glassy, et al., 2011), and registers essential metadata and persistent identifiers in library catalogues such as TIBORDER (http://tiborder.gbv.de) and GetInfo (https://getinfo.de). This procedure corresponds to the recommendation of the STD-DOI project and its successor project KOMFOR (http://www.komfor.net), which grant scientific primary data an identity of their own.

With respect to persistent identifier registration, the WDC Climate acts as a publication agency, which is responsible for precursory quality assessment and long-term storage after the data have been published. The unique identifiers used in this context are Digital Object Identifiers (DOI), such as "10.1594/WDCC/CLM_A1B_2_D3", and Uniform Resource Names (URN), such as "urn:nbn:de:tib-10.1594/WDCC/CLM_A1B_2_D30". Scientists can use both identifiers to directly reference the metadata and the primary data in their articles. Readers of these articles can then, for example, enter the DOI on the DOI Foundation resolver homepage (http://dx.doi.org) and will be directed to the current "landing page". The WDC Climate provides a "Compact Page" as the landing page, which contains the full set of metadata and, of course, downloads links to the primary data.

Up to the year 2009, the WDC Climate registered DOI and URN by calling web services of the German National Library of Science and Technology (TIB). But in 2009 several international institutes joined together and founded the DataCite consortium to establish a global organisation for registering DOIs for research data (Brase, 2009). DataCite is a registration agency of the IDF (International DOI Foundation) (http://www.doi.org) and provides its own technical interfaces for DOI registration. Accordingly, the WDC Climate switched to the web services of DataCite.

As publication renders the data more visible, quality becomes an even more crucial factor. To ensure the high quality of the primary data and metadata, the WDC Climate set up an appropriate procedure, which comprises a review of both data types by a scientist familiar with the data (Scientific Quality Assurance, cf. Subsection 3.1.3) and a review of the technical aspects of the work, performed by a data publication agent at WDC Climate (Technical Quality Assurance, see upcoming Subsection 3.2.4).

The procedure, which had been executed before the activities of the research project started, involved several manual steps, hardly any comfortable user interfaces (the user had to provide an XML file including all the metadata), and involved a significant amount of communication between the scientist and the publication agent. Moreover, it was not possible to describe observation instruments or platforms or to enter details of the quality assurance measures performed.

In the course of the research project, the data publishing experts and the computer scientists first modelled the publication process using the Business Process Model and Notation (Grosskopf, Decker, & Weske, 2009). Then they discussed the scope for improving the process and updated the model accordingly. Besides the general process workflow, the project members defined a comprehensive metadata set and drafted forms used to conveniently enter these details. Both the final process model and the input form drafts were essential prerequisites for developing the Atarrabi publication software (see Subsection 4.1).

The WDC Climate is also involved in the CMIP5 project (see Subsection 3.1.4). Besides setting the standards for scientific quality, it defines the formal and technical requirements for data publication (Stockhause, Höck, Toussaint, & Lautenschlager, 2012). As an important part of the quality assurance process, the working group responsible defined a review process by the scientist, which ends with the scientist's approval of the data and metadata. Since this procedure and communication between the scientist and the publication agent are among the main functions of Atarrabi, the decision was made to use Atarrabi for the review process of CMIP5 data. Atarrabi originally did not cover simulation data and had to be extended to support this data type.

3.2.4 Technical Quality Assurance

Scientific Quality Assurance (cf. Subsection 3.1.3) deals with scientific aspects of the data, such as badly chosen or broken measurement devices and implausible or unphysical values. This kind of quality assurance has to be performed by a scientist familiar with meteorological experiments and the data generated. Technical Quality Assurance, however, covers technical aspects affecting data management and thus has to be conducted by a data management specialist.

Routines run in this context are aimed at finding out, for example, whether all relevant datasets of primary data are actually available and publicly accessible in CERA, whether they contain any data at all, and whether they are correctly linked to the experiment of interest. The size values of the datasets are verified and aggregated into dataset group or experiment level, in order to define a granularity level for climate model data that fits in with the granularity of reference lists in scientific literature. These values are necessary for registration at DataCite and are used as a check sum to validate the data. Moreover, the primary data and metadata are checked for consistency, e.g., with respect to time intervals and variable names.

To collect the findings of the TQA procedure, the software needs to provide a form to enter TQA data. Atarrabi implemented this form as part of the Agent Wizard (see Subsection 4.3.2).

3.3 The Process Automation Perspective

Computer scientists made a vital contribution to the project from the point of view of process automation. At first, they analysed the publication process in cooperation with the data publication experts at WDC Climate and created a business process model. Based on this model and other requirements, they designed and implemented the Atarrabi publication software. Section 4 will present selected requirements dealt with during software development and introduce the actual implementation.

4 THE ATARRABI PUBLICATION SYSTEM

As described in Subsection 2.2 and in the preceding section, the web-based software solution Atarrabi has been developed to control and support the data publication procedure according to the newly defined and optimised process. Atarrabi covers the workflow beginning with the scientist's SQA on metadata and closing with DOI and URN registration at DataCite (to release an experiment beforehand and to denote the person conducting the publication process, a dataset is merely added to a dedicated table in the CERA database).

In the following, the main functionalities and characteristics of the final product will be described in detail (please see Figure 3 for review of the publication procedure). For scientists using Atarrabi for the first time, a quick-start tutorial can be downloaded from the Atarrabi homepage (http://cera-www.dkrz.de/atarrabi) at WDC Climate.
The Atarrabi system has been successfully deployed in the WDC Climate infrastructure (http://cera-www.dkrz.de/atarrabi). The source code of Atarrabi has been made available under a GNU General Public License on SourceForge (http://sourceforge.net/projects/atarrabi).

4.1 Requirements and Software Design

As a very basic requirement, the software had to be web-based, and it also had to be capable of being integrated into the existing IT infrastructure at WDC Climate. This required the software to be able to use an existing LDAP-server for user authentication and to be deployable on an Oracle WebLogic application server. The user interface had to be provided in English, but it should be easy to make it available in other languages as well. Metadata edited in Atarrabi had to be downloadable as a PDF file.

The publication process starts at a point when the primary data and metadata have already been stored in the long-term archive. In the course of the process, the metadata have to be displayed, changed, and augmented in Atarrabi. This calls for a data exchange interface between Atarrabi and the CERA database. Since the CERA database is a production system that is widely used, Atarrabi must not have direct writing access to the main data tables. Moreover, to keep the data basis for the scientist as stable as possible, it was decided that once the scientist's review process for an experiment has started and the existing data have been imported into Atarrabi, there would be no further synchronisation with CERA for this process instance.

The process model created had to be integrated into the software to control the actual workflow. This is why it was highly desired that the software framework support a business process engine of some kind. Using a business process engine allows for easy definition of and modifications to workflows. Moreover, these engines implement common business process management concepts, such as task lists, user roles, and conditional workflows.

The main functional areas of Atarrabi include the user interfaces provided to the scientist and the publication agent for the purpose of editing metadata. There are a vast number of metadata fields, and these should not be displayed on a single page. In collaboration with the meteorologists and data publication experts, the metadata fields have been logically grouped and spread across multiple pages. These pages can be traversed the same way as in a software installation wizard and are, hence, called the "Scientist Wizard" and the "Agent Wizard". In the Scientist Wizard, exhaustive help texts have to be provided to support first-time users. In both the Scientist Wizard and the Agent Wizard, scientists and publication agents must be able to leave relevant text messages to each other. This avoids time-consuming email correspondence or agreements made orally.

To design the user interface and the data fields displayed on screen, the computer scientists created virtual windows as described by Lauesen (2005). This approach uses very rough sketches of the application views to discuss the contents and layout. Once all parties had agreed to such a draft, the view designers and back-end programmers used it as a requirement document. Figure 5 shows an example of a virtual window used in the project.

As part of the data publication process, the scientist has to define the data authors, the DOI contact person, and the contributing institutions. At this point, Atarrabi has to provide a form to search for the person and institute entries already stored in CERA. If the person or institute cannot be found, the scientist can create a new entry without any difficulty. A similar approach had to be implemented for metadata fields displayed as dropdown lists (e.g., units, relation types, languages, etc.). These lists of values are provided by the CERA database and have to be automatically imported into Atarrabi at the first startup and at any desired point in time.

Atarrabi has to provide different user experiences for scientists and publication agents. The scientist must be able to initiate a new publication process; he must also be capable of continuing and tracking process instances he or she has started before. For this function, the developers have created the "My Publications" area, which shows the process instances performed by the scientist. The area dedicated to the publication agent has been named "Agent Workspace" and provides such functions as assigning a process instance to the agent, starting the agent review for a submitted publication instance, and finally, registering the DOI. Atarrabi must provide forms and links for feedback of any kind to all users.

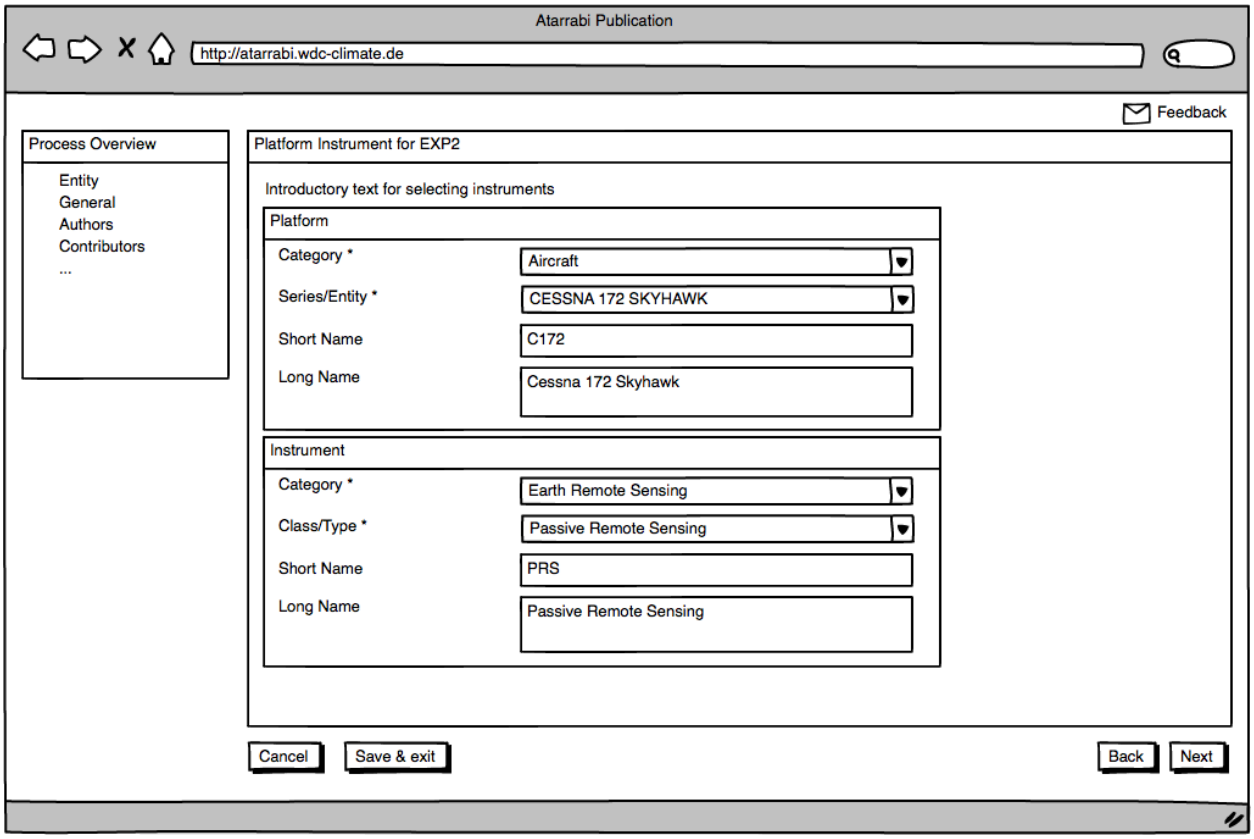

Figure 5. A virtual window used in the project for designing the Scientist Wizard

4.2 Implementation of the Atarrabi Workflow System

To find appropriate software technology facilitating the requirements mentioned above, the computer scientists performed an extensive technology evaluation. Finally, the Seam Framework (http://www.seamframework.org) by JBoss was selected as a basis because it is a Java-based, open-source web framework that integrates the jBPM workflow engine (http://www.jboss.org/jbpm), which was supposed to support the publication workflow. Seam supports the generation of PDF files and can use appealing user interface libraries such as JBoss RichFaces and jQuery, which are compatible with numerous browsers. In addition, Seam-based applications can be deployed on WebLogic application servers. In legacy web-based applications, users are not informed about the outcome of an action until a new page is rendered, which may lead to long-winded and cumbersome user interactions. Seam uses modern technologies (e.g., AJAX) to provide a responsive user interface that immediately shows the outcome of an action (e.g., instant removal of a data row after clicking a delete icon).

4.3 The Wizards of Atarrabi

4.3.1 The Scientist Wizard

After the scientist has logged in, he or she is guided to the "My Publications" page. On this page he or she can start a new publication process by selecting an experiment to publish and will be directed to the first page of the "Scientist Wizard" (see Figure 6). This wizard supports the scientist in performing the SQA of the metadata and in adding details of the quality assurance measures applied to the primary data. To keep the displays as simple and clear as possible, the metadata fields have been divided into several logical sections, e.g., general metadata, authors, relations. The scientist can move from one section to the next by traversing the wizard. Such a principle is widely known in the area of software installation wizards, but here, in addition, the user can leave the wizard at any time and finish it at a later date without losing the values entered.

All wizard pages provide exhaustive help texts. Where appropriate, the system offers special input fields (e.g., a calendar for date selection, dropdown menus for selecting an entry from a list of given values, a map that visualises the coordinates for spatial coverage). If persons or institutes have to be specified, the system enables the user to search the entries already stored in CERA. If the person or institute is not listed, the scientist can create a new entry. It has been specified whether a field is mandatory or not, and all fields are validated if the user attempts to proceed to the next wizard page.

Particularly noteworthy is the wizard page for specifying the scientific data quality procedure (see Figure 7). On this page, the author can summarise the quality checks performed on the primary data. Quality information exceeding this summary (check procedure details, protocols, images) can be provided as additional files that are uploaded to the database.

On the last page, the scientist is presented with a summary containing all the data he or she has entered while traversing the wizard. It is possible to download this summary as a PDF file. If all data are correct and complete, the scientist submits the data and triggers the next process step, the review by a publication agent. To send questions or notes to the publication agent, each wizard page provides a dedicated text input area. These comments will be visible to the publication agent as soon as the scientist finishes the wizard.

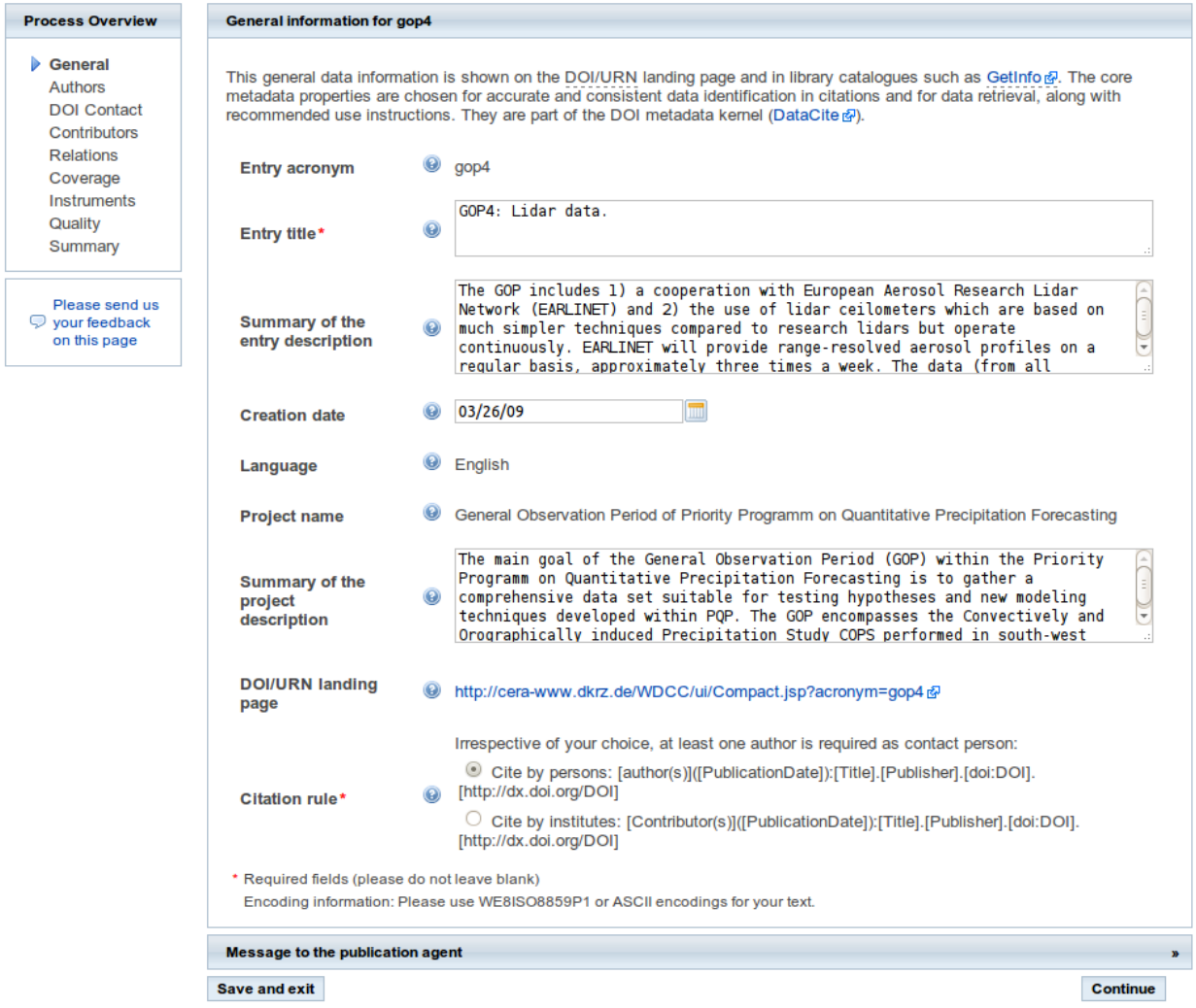

Figure 6. The first page of the Scientist Wizard

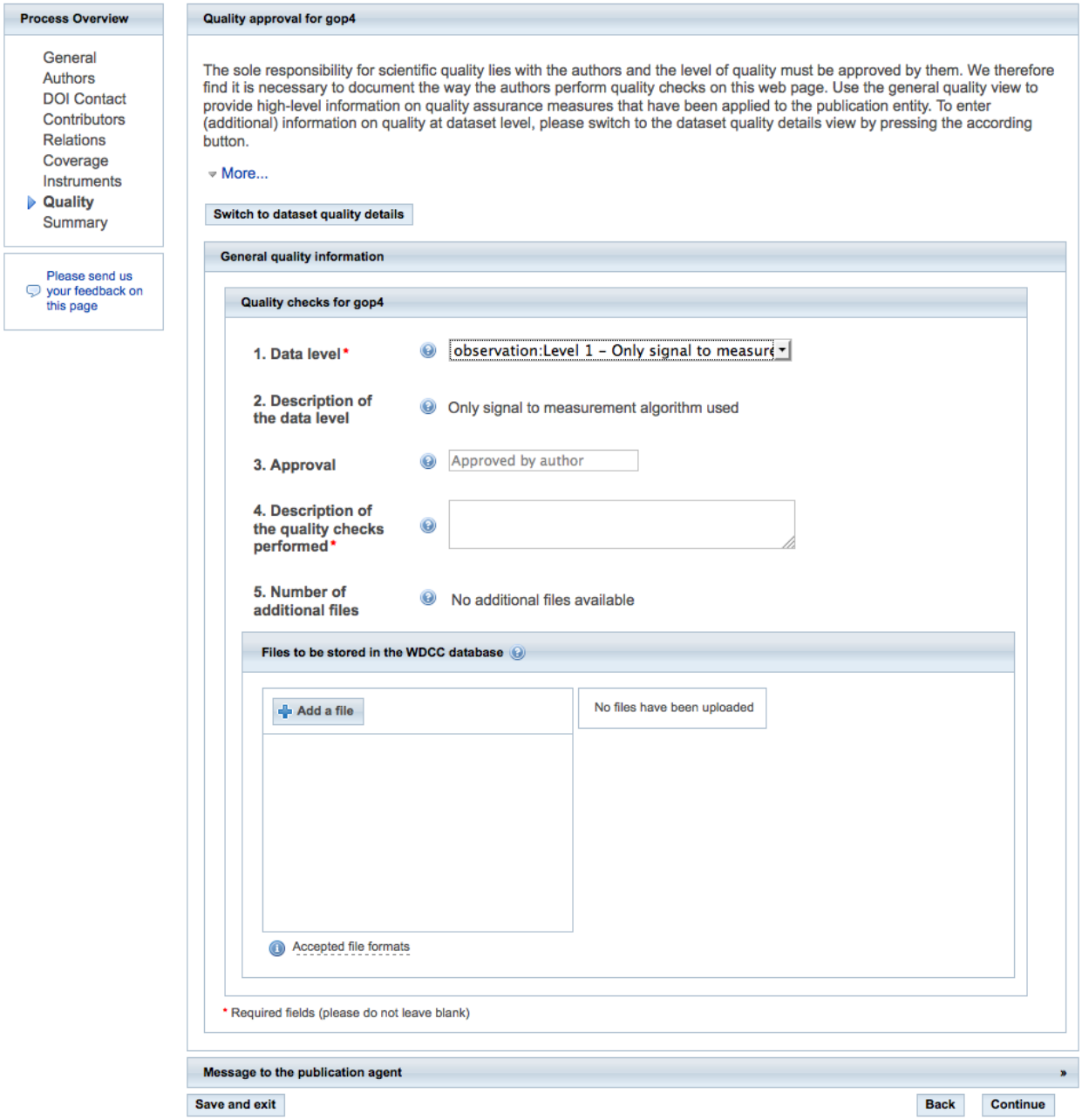

Figure 7. The Wizard page for the specification of the scientific data quality procedure

4.3.2 The Agent Wizard

As soon as a scientist has completed the Scientist Wizard, Atarrabi sends an email to all publication agents at WDC Climate to inform them that a new publication entity needs reviewing. After login, agents have access to the "Agent Workspace" page. This page shows all process instances whose Scientist Wizards have been finished and have to be assigned to an agent. By clicking the appropriate link, the agent can move one of these process instances to his or her personal task list and begin the review process by starting the "Agent Wizard". The Agent Wizard is similar to the Scientist Wizard and enables the agent to review and change the metadata. The main differences in comparison to the Scientist Wizard are as follows:

- Most of the Agent Wizard pages show the metadata fields in a two-column view. While the column on the left-hand side shows the data submitted by the scientist, the column on the right-hand side displays the data

currently stored in CERA. If the data differs between the columns, the row is highlighted using a yellow background colour. In this way the agent is easily attracted to the values modified (see Figure 8).

- If the scientist has written a note to the publication agent, this message is displayed in a text box right at the top of the corresponding wizard page. To answer these messages or to add other comments, the publication agent has been provided with a text field as well. In case the agent rejects the data submitted, these messages will be visible to the scientist in the Scientist Wizard when it is re-visited
- There is an additional wizard page for documenting the TQA checks (see Subsection 3.2.4).
- Having reached the last page of the Agent Wizard, the agent has to decide whether he or she accepts or rejects the data submitted. In the former case, the process continues with the identifier registration step; in the latter case, the process instance is assigned back to the scientist who has to re-enter the Scientist Wizard and re-submit the data after having corrected it.
- The publication agent will see the messages from the scientist as a consolidated list, right at the beginning of the Agent Wizard. Provided the agent does not accept the data submitted, the scientist will have to re-enter the Scientist Wizard and will see the agent's messages at the top of the display concerned.

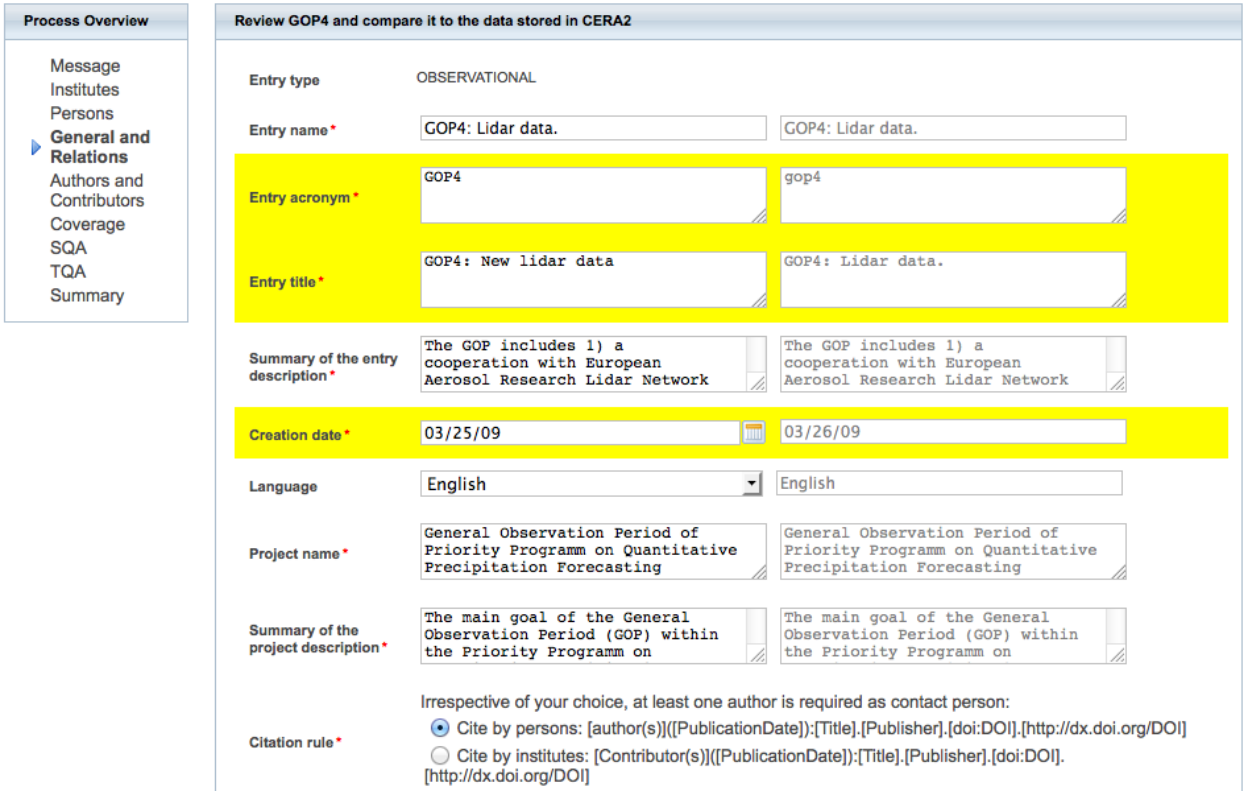

Figure 8. Highlighting of changes in a view of the Agent

4.4 Data Import and Export

There are a variety of situations in which Atarrabi relies on external data. Examples include the list of experiments available for publication, the metadata that have already been stored in the long-term archive to initialise the input fields in the wizards, the list of values for units, instruments, relation types, etc. Considering the project objective that the software should be easily adaptable to other data centres (cf. Subsection 4.5), the developers have defined an interface of read-and-write operations that have to be implemented by the external database. For the CERA database at WDC Climate, these interface operations were implemented as PL/SQL procedures (a programming language for Oracle databases). In the end, Atarrabi just has to call procedures such as "getUnits" to make CERA return a table of data as specified in the interface.

The most important advantage of this approach over using an ordinary database connection and sending SQL statements is the fact that Atarrabi does not depend on the external database schema. As a consequence, Atarrabi does not have to be adapted if there are any modifications to the database schema since only the procedures stored in CERA have to be updated.

The experiment metadata changed and added in Atarrabi have to be committed back to the external database at some point. For this purpose, Atarrabi generates an XML file that strictly follows a specified schema and contains all metadata information gathered in the wizards. This file is sent over to CERA, where the data is extracted and the datasets are updated accordingly in a special routine procedure. In this way an expert from the data centre can supervise all changes relevant to the database if required, and again, Atarrabi does not depend on the database schema. To provide agents and users with a document that is easy to read and printer-friendly, Atarrabi can export the experiment metadata as a PDF file.

Some preliminary measures have to be taken to register a DOI and a URN. First of all, Atarrabi exports an XML file containing all the metadata entered into the CERA database. Based on this data, CERA is updated to the new values. After that, a further XML file is generated in a routine procedure in the CERA database. The latter contains an exactly specified subset of the metadata and is compliant with the DataCite Metadata XML schema (http://schema.datacite.org). This XML file also contains the actual DOI and URN, which have been generated by CERA based on the experiment acronym. The agent can view this XML file in Atarrabi to perform a final quality assurance test. Finally, Atarrabi calls a web service provided by DataCite to transmit the DataCite XML file. Once this call has been completed successfully, the persistent identifiers are registered and the scientist is informed immediately by email.

4.5 Adaptability to other Environments

The main objective of the project has been to design a process and a software that can be adapted to the needs of other environmental sciences and long-term archives. In the end, the system has to be flexible with respect to the following aspects:

Data types
Originally, the project was aimed at supporting data that were produced as a result of observation experiments. During the project period, the support of model data produced by climate simulations in CMIP5 projects became an additional requirement. In fact, the metadata of observational data and model data share most of the metadata fields but differ in some details. Metadata for observational data, for example, have extra fields containing instrument and platform information while CMIP5 metadata have an additional title field and a check box that indicates that the scientist has filled in the CMIP5 questionnaire. Beside additional fields, different data types can implicate additional or reorganised wizard pages containing these fields. Flexibility with respect to different data types has been successfully demonstrated by the fact that Atarrabi supports the CMIP5 data type in addition the observational data type.

Data sources
As explained earlier, there is no access to external data sources that depend on a certain database model. For importing data from the long-term archive CERA, Atarrabi employs stored procedure calls. In turn, XML files are used to export data from Atarrabi back to CERA. In this way, Atarrabi can rely on a very stable access interface. An additional database can be used as a data source if it implements the defined interface.

Process flows
The publication process may differ from one registration agency to the next. Thanks to the flexible jBPM workflow engine, it is quite easy to vary the main process flow. This includes the interplay of the scientist, the publication agent, and the registration agency. jBPM even allows for the parallel execution of process instances that are based on differing versions of the process description (e.g., to finish older process instances that have not yet been completed).

Interface customisation
Changes to the layout and text contents can easily be implemented using common web technologies, such as cascading style sheets, HTML, and embedded pictures (e.g., a logo of the data centre). Moreover, the Seam framework provides comprehensive internationalisation features that add support for additional languages.

5 CHALLENGES AND FUTURE WORK

Atarrabi has been designed to be easily adaptable to other environments. Accordingly, it should be easy to tailor Atarrabi to the needs of other data repositories in the field of environmental sciences. More extensive adaptations might make Atarrabi applicable to institutional repositories that deal with data of various disciplines or qualify Atarrabi as an easy-to-use data preparation tool provided by data registration agencies such as DataCite.

Applying quality assurance methods known from traditional text publications to data publication can be extremely difficult and is the subject of ongoing research. Performing an effective peer-review of data is challenging because of the variety of file formats as well as the vast amounts of data, and solutions have to emerge from scientific practice. A possible approach might involve the provision of standardised software tools which automate statistical quality assurance tests on general data. Independent reviewers can then use the output of these tools to assess the quality of the datasets and decide, just as in traditional publications, whether the data must be revised or can be published. Until these tools have been established, a pragmatic approach might involve the implementation of a user interface to annotate data after they have been reused and after quality or other relevant issues have been addressed. It goes without saying that this type of feedback information should be integrated into platforms where scientists usually search for and download data.

6 CONCLUSION

The web-based Atarrabi platform for the preparation and publication of environmental data was successfully introduced at WDC Climate and has been in use since 2010. Atarrabi supports the publication process by presenting the different user types with dedicated, easy-to-use interfaces and by guiding the scientist and the publication agent through the review process. For this purpose, the metadata have been logically grouped and spread over several pages that form the Scientist and the Agent Wizard. User feedback on this approach has been very positive. Scientific and technical quality assurance and the relevant documentation play an essential role in the publication and reuse of data. In the course of the project, the R-package “qat” was developed, which supports the scientist in finding irregularities in primary data. Atarrabi provides data entry forms to enter details of the quality assurance measures performed, and it also offers an interface to upload the relevant files.

Observational data published with Atarrabi are COPS (Convective and Orographically-induced Precipitation Study) data, field data of the Meteorological Institute of University of Hamburg, and glacier mass balance data of Austria. Moreover Atarrabi has been successfully used in the global climate-modelling project CMIP5. All published data can be found via the WDC Climate web portal (http://cera-www.dkrz.de/WDCC/ui/FindDoiPublications.jsp).

7 REFERENCES

Brase J. (2009) DataCite - A Global Registration Agency for Research Data. *Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, COINFO '09*, pp. 257-261.

Brase, J. (2004) Using Digital Library Techniques – Registration of Scientific Primary Data. *Lecture Notes in Computer Science 3232/2004*, pp. 488-494. Retrieved April 25, 2012 from the World Wide Web: http://dx.doi.org/10.1007/978-3-540-30230-8_44

Committee to Review the Intergovernmental Panel on Climate Change (2010) Climate Change Assessments, Review of the processes and procedures of the IPCC. Retrieved April 25, 2012 from the World Wide Web: http://reviewipcc.interacademycouncil.net/report.html

Costello, M.J. (2009) Motivating Online Publication of Data. *BioScience 59* (5), pp. 418-427. Retrieved April 25, 2012 from the World Wide Web: http://dx.doi.org/10.1525/bio.2009.59.5.9

Duerr, R.E., Downs, R.R., Tilmes, C., Barkstrom, B., Lenhardt, W.C., Glassy, J., Bermudez, L.E., & Slaughter, P. (2011) On the utility of identification schemes for digital earth science data: an assessment and recommendations. *Earth Science Informatics 4*, pp. 139-160. Retrieved April 25, 2012 from the World Wide Web: http://dx.doi.org/10.1007/s12145-011-0083-6

Düsterhus A. & Hense. A. (2012) Advanced information criterion for environmental data quality assurance. *Advances in Science and Research 8*, pp. 99-104.

Eaton, B., Gregory, J., Drach, B., Taylor, K. & Hankin, S. (2011) NetCDF Climate and Forecast (CF) Metadata Conventions. Retrieved April 25, 2012 from the World Wide Web: http://cf-pcmdi.llnl.gov/documents/cf-conventions/1.6/cf-conventions.html

Toussaint, F., Lautenschlager, M. & Luthardt, H. (2007) World Data Center for Climate Data – Support for the CEOP Project in Terms of Model Output. *Journal of the Meteorological Society of Japan (85A)*, pp. 475-485.

Grosskopf, A., Decker, G. & Weske, M. (2009) *The Process: Business Process Modeling Using BPMN*. Tampa, FL: Meghan-Kiffer Press

Guilyardi, E., Balaji, V., Callaghan, S., DeLuca, C., Devine, G., Denvil, S., Ford, R., Pascoe, C., Lautenschlager, M., Lawrence, B., Steenman-Clark, L., Valcke, S. (2011) The CMIP5 model and simulation documentation: a new standard for climate modelling metadata. *CLIVAR Exchanges (Special Issue 56) 16* (2), pp. 42-46. Retrieved April 25, 2012 from the World Wide Web: http://www.clivar.org/sites/default/files/imported/publications/exchanges/Exchanges_56.pdf

Gura, T. (2002) Scientific publishing: Peer review, unmasked. *Nature 416*, pp. 258-260.

Henneken, E.A. & Accomazzi, A. (2011) Linking to Data – Effect on Citation Rates in Astronomy. Retrieved March 7, 2012 from the World Wide Web: http://arxiv.org/abs/1111.3618v1

Hense, A. & Quadt, F. (2011) Acquiring High Quality Research Data. *D-Lib Magazine 17* (1/2). Retrieved April 25, 2012 from the World Wide Web: http://dx.doi.org/10.1045/january2011-hense

Lauesen, S. (2005) *User Interface Design: A Software Engineering Perspective*, Harlow: Addison Wesley

Lautenschlager, M. (2011) Institutionalisierte "Data Curation Services". In Büttner, S., Hobohm, H.-C., Müller, L., (Eds.), *Handbuch Forschungsdatenmanagement*, Bad Honnef: Bock + Herchen Verlag

Lautenschlager, M., Toussaint, F., Thiemann, H. & Reinke, M. (1998) The CERA-2 Data Model. Retrieved April 25, 2012 from the World Wide Web: http://dx.doi.org/10.2312/WDCC/DKRZ_Report_No15

Lawrence, B., Jones C., Matthews, B., Pepler, S., Callaghan S. (2011) Citation and Peer Review of Data: Moving Towards Formal Data Publication. *The International Journal of Digital Curation 6*, pp. 4-37.

Lorenz, E. (1963) Deterministic Nonperiodic Flow, *Journal of the Atmospheric Sciences 20*, pp. 130-141.

Meehl, G.A., Covey, C., Taylor, K.E., Delworth, T., Stouffer, R.J., Latif, M. et al. (2007) The WCRP CMIP3 Multimodel Dataset: A new era in climate change research. *Bulletin of the American Meteorological Society 88*, pp. 1383-1394.

Palmer, T.N. (2012) Towards the probabilistic Earth-system simulator: a vision for the future of climate and weather prediction. *Quarterly Journal of the Royal Meteorological Society 138*, pp. 841-861.

Parsons, M.A., Duerr, R. & Minster, J.-B. (2010) Data Citation and Peer Review. *Eos, Transactions, American Geophysical Union 91 (34)*, pp. 297-299.

Paskin, N. (2005) Digital Object Identifiers for scientific data. *Data Science Journal 4*, pp. 12-20. Retrieved April 25, 2012 from the World Wide Web: http://dx.doi.org/10.2481/dsj.4.12

Pfeiffenberger, H. & Carlson, D. (2011) "Earth System Science Data" (ESSD) – A Peer Reviewed Journal for Publication of Data. *D-Lib Magazine 17* (1/2). Retrieved April 25, 2012 from the World Wide Web: http://dx.doi.org/10.1045/january2011-pfeiffenberger

Pöschl, U. (2010) Interactive open access publishing and public peer review: The effectiveness of transparency and self-regulation in scientific quality assurance. *IFLA Journal 36*, pp. 40-46.

Science Staff (2011) Dealing with Data – Introduction: Challenges and Opportunities, *Science 331 (6018)*, pp. 692-693. Retrieved from the World Wide Web: http://dx.doi.org/10.1126/science.331.6018.692

Stockhause, M., Höck, H., Toussaint, F. & Lautenschlager, M. (2012) Quality assessment concept of the World Data Center for Climate and its application to CMIP5 data. Geoscientific Model Development Discussions, 5, 781-802. Retrieved June 25, 2012 from the World Wide Web: http://dx.doi.org/10.5194/gmdd-5-781-2012

(Article history: Received 16 July 2012, Accepted 19 October 2012, Available online 3 November 2012)