Assessment Information

CoreTrustSeal Requirements 2017–2019

Repository:	PANGAEA - Data Publisher for Earth and Environmental Sciences
Website:	https://www.pangaea.de
Certification Date:	17 June 2019
This repository is owned by:	**MARUM, University Bremen / Alfred Wegener Institute, Bremerhaven**

PANGAEA - Data Publisher for Earth and Environmental Sciences

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository, Research project repository

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

The information system PANGAEA is operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth system research. PANGAEA is hosting a wide range of domains (trans-disciplinary) - compare the PANGAEA home page (https://www.pangaea.de/). Main disciplines covered are biogeochemistry, oceanography, geological, atmospheric, and biological sciences. PANGAEA currently keeps >14 billion data items (numeric, textual, and binaries) comprised in >380.000 data sets. Up to now data management has been supplied for >300 national to international projects (www.pangaea.de/projects/). PANGAEA is hosted by the Alfred Wegener Institute, Helmholtz Center for Polar and Marine Research (AWI - www.awi.de) and the Center for Marine Environmental Sciences, University of Bremen (MARUM - www.marum.de).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository’s Designated Community.

For data archiving and publication as well as re-usage of data: researchers worldwide including those from our host institutions working in the domains covered by the editorial of PANGAEA. Particular emphasis is given to data management of any national, EU, and International Science Projects. PANGAEA data curators are working closely together with scientists (data management as an integral part of science). Compare the PANGAEA project listing: https://www.pangaea.de/projects/

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

D. Data-level curation – as in C above; but with additional editing of deposited data for accuracy

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

Curation ensures completeness, correctness, and interoperability (machine readibility) of data and metadata. In average 4 working hours by domain experts are needed for each data submission. Curation is supported by an editorial system ensuring structural and semantic harmonization of data.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Outsource Partners. If applicable, please list them.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

PANGAEA is operating the "World Radiation Monitoring Center" (WRMC - http://bsrn.awi.de/) and as such is accredited as a "Data Collection and Processing Center" (DCPC) in the "World Meteorological Organisation Information System" (WIS).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

ORGANIZATIONAL INFRASTRUCTURE

I. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The information system PANGAEA is operated as an Open Access library aimed at archiving, publishing and distributing georeferenced data from earth and environmental sciences. It focuses on georeferenced observational data, experimental data, and models/simulations. Citability, comprehensive metadata descriptions, interoperabillity of data and metadata, a high degree of structural and semantic harmonization of the data inventory as well as the committment of the hosting institutions ensures the long-term usability of archived data. The general IT and technological developments in particular with regard to measurement and device technology led to an exponential growth of data production and opened up new ways for large scale and interdisciplinary research approaches. There is a trend towards open networked data infrastructures and „Data Intensive Science“. To effectively meet the resulting challenges and to use the new opportunities in an optimal manner is the general aim of PANGAEA. The intense linkage with the wide spectrum of science and technology activities in our host institutions thereby offers a helpful background and particular motivation. The matching demands create synergies which can be used in many ways (cmp. also the AWI Data Flow Framework: https://www.awi.de/en/about-us/service/computing-centre/data-flow-framework.html).

On the one hand PANGAEA shall give targeted support for research in the host institutions, on the other hand the information system shall also render a service to the wider communities. The latter not only adds to the reputation of host institutions but is also a vital contribution to the globally evolving research infrastructures and to research in general.

The role of PANGAEA for the host institutions is referred to on their websites. For the AWI: https://www.awi.de/en/about-us/service/computing-centre/data-products.html - and MARUM: https://www.marum.de/en/Infrastructure/PANGAEA-Data-Publisher-for-Earth-and-Environmental-Science.html

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

II. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

PANGAEA uses Creative Commons (https://creativecommons.org/) licenses, mostly CC-BY, in a few cases CC-BY-NC (geophysics - not recommended anymore). For software (few data sets) GPL (https://www.gnu.org/licenses/licenses.en.html) is used. Data related to genetic resources are handled according to the Nagoya Protocol (https://www.cbd.int/abs/about/default.shtml/). A minor part of the data (<1%) are under moratorium, mostly data from running scientific projects: protection is applied on individual and group level, metadata are public. We have recently moved to CC version 4.0.
Information with recommendations to choose a suitable CC license for data submittors can be found here: https://wiki.pangaea.de/wiki/License. Data ownership/copyright stays at original submittor (no copyright transfer). License changes require confirmation by data owner. Recommendation by PANGAEA is to use CC-BY, but also CC0 is possible. In general all CC licenses are allowed, but NC and ND licenses are discouraged as they limit usage. In any case data providers have to decide for one of the supplied options.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
It is recommend to develop a breach policy in the future.

III. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

PANGAEA operates since 25 years. Being part of the research infrastructure, the host institutions ensure long term support for storage and dissemination of data. As part of the AWI MARUM cooperation agreement (AMAR) of 2014 both institutions state their collaboration with respect to PANGAEA. Since March 2011 PANGAEA is controlled by a steering committee, which has set up a coordination unit. The latter organizes technical operations and the further development of PANGAEA as well as the data contents and the editorial work. Other tasks include the staff and financial planning and management of third party funded projects. The work of the coordination unit is supervised by the steering committee. This committee - composed of representatives from AWI and MARUM including the directors - meets regularly to discuss relevant topics, opportunities, and risks and takes strategic decisions to ensure the sustainability and further development of PANGAEA.

Beyond the host institutions: PANGAEA is well embedded into the national and EU research funding landscape and gets continuous and substantial support through project grants. Within the emerging German National Research Data Infrastructure (NFDI) PANGAEA plays a significant role.

Since 2014 the PANGAEA technical and organisational structures as well as our activities are guided by a strategy plan, which is currently under preparation for the next period until 2023 (https://goo.gl/jm5Nk5)

PANGAEA is the primary custodian for all archived data.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IV. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

All data editors affiliated with PANGAEA are familiar with the implications of the rules of good scientific practice for the management of scientific data. They are domain experts with knowledge about the requirements in the specific science fields.
During the data submission process, users are made aware of their responsibility for the correctness of data and metadata (see https://pangaea.de/submit/). Treatment of sensible or restricted data is negotiated in direct communication with data providers. If necessary, data can be password secured (individual or group level). In such cases the principle investigator is responsible for giving access rights.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

V. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

PANGAEA is long term operated by its hosting institutions, the Alfred Wegener Institute for Polar and Marine Research (AWI) and the Center for Marine Environmental Sciences (MARUM) at the University Bremen (see http://wiki.pangaea.de/wiki/File:AWI_MARUM_PANGAEA.pdf). The annual overturn of AWI and MARUM together is around 140 Mio Euro. The annual overturn of PANGAEA is around 1 Mio Euro. Basic costs are covered by the host institutions. Funds for data management are mostly received from science projects (international, EU, national). Staff responsible for data management (data editors ~ 15 employees) are well trained and most of them are experts in their corresponding fields. Technical staff (~ 5 employees) develops and maintains specific software and services for PANGAEA. The AWI computer center (~50 employees) supplies basic IT services. Working hours are covering in average 12 hours a day, in case of system failures there is also limited support at night times and on weekends.
Staffs affiliated with PANGAEA meets regularly every 6 weeks to exchange knowledge, resolve issues, and discuss the status and development of the system. New staffs get special training. Beyond all staffs use their educational leave (10 days/year) for further training and the acquisition of new skills. Staffs are recruited mostly from the academic environment of AWI, MARUM, and further institutes related to earth and environmental sciences.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VI. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

PANGAEA is controlled by a steering committee (SC) composed of representatives/scientists from AWI and MARUM including the directors. The SC meets regularly to discuss relevant topics, opportunities, and risks and takes strategic decisions to ensure the sustainability and further development of PANGAEA.
Beyond the host institutions PANGAEA on the national level is part of the German Federation for Biological Data - GFBio (https://www.gfbio.org/) and the German Network for Bioinformatics Infrastructure – de.NBI (https://www.denbi.de/). On the European level PANGAEA is participating or has been participated in some long-term infrastructure projects (ESFRI - e.g. ENVRI+, EMSO, Elixir). In addition PANGAEA is very active in the Research Data Alliance thus learning from other organizations and profiting from the overall development in research data infrastructures.
As a funded partner PANGAEA has supplied data management services for numerous national to international projects (www.pangaea.de/projects) leading to intensive collaborations with scientists. At an informal level PANGAEA staff keeps linkages with many scientists within the host institutions and worldwide discussing and giving advice for handling of new data types, methodologies, and the processing of data etc.
A large part of the communication with users is covered by our ticketing system (data submission, contact form) allowing to assemble information relevant for the further development of PANGAEA in an organized and documented way. As a further communication channel we have started in 2019 a twitter account (https://twitter.com/pangaeadatapubl).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

DIGITAL OBJECT MANAGEMENT

VII. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Data entities are fully documented (conform ISO19115, DIF, DC, compatible to ABCD, Darwin Core) and archived into a relational database (RDB). A copy of each data set (with checksum and timestamp) is marshalled to disk as tab delimited files. Any changes to data and metadata are recorded in the editorial system (update history and versioning). In case of losses or accidental changes to data and metadata versions in the RDB can be compared with copies on disk. The data submission system (https://www.pangaea.de/submit/) keeps the original files uploaded by data providers. Data ingest includes V&V of supplied data and metadata. Manual checks include e.g. validity of used observation types and methods or the identity checks of data providers. PANGAEA is requesting and checking ORCIDs for all authors (registration includes request of an ORCID), nevertheless, identity of depositors is currently limited to checking institutional affiliations and mail addresses manually. Useful in this context is the correspondence with all data providers via our ticketing system. Automatic routines include correctness of data values (data type, precision), identification of outliers and range checks. Access to data is always via the metadata catalogue. Data sets are always delivered as complete entities with metadata included. In case data from different data sets are compiled from the PANGAEA data warehouse the result matrix always includes the DOI for each data item thus ensuring identification of the original data sets. Principle investigators/authors as well as any institutution or laboratory/facilities/methodologies/events contributing to the production or processing of data are recorded in the metadata (provenance). With a few exceptions data sets are distributed with Creative Commons Attribution licence (http://creativecommons.org/licenses/by/3.0/) ensuring that authors of the data are cited.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

VIII. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

PANGAEA accepts data from all fields of earth and environmental sciences (compare www.pangaea.de/submit). Data can be of any type (numerical, text, or binary). For data submissions an issue tracking system is supplied (https://issues.pangaea.de/), which allows data providers to specify initial metadata and upload of data files. All subsequent communication between data editors and data providers is done via the issue tracking system thus documenting each data submission (SIP). Each data submission is thus guided by domain experts with the aim to optain complete and correct data and metadata. Submitted data files, which can be of any readable format, are transformed to the PANGAEA internal import format. Metadata are transferred into the PANGAEA editorial system whereby the completeness and correctness of information is checked. Data editors also check the validity of used methods, whether units are consistent with usual standards, and whether the data set as a whole complies to possible community standards. The relational database underlying the editorial system ensures that supplied metadata are consistent with the existing information inventory. Usage of the terminologies (like WoRMS or Chebi) built into the editorial system further contribute to the harmonization of data and metadata. Numerical data are checked for outliers, range of values, and correct geocoding (deviating values are flagged). Nominal data are checked for consistency with scales supplied. Binary data are tested with customary software. Some high volume data products are furnished with thumbnail information (previews). All transformations on data and metadata are negotiated with the data providers. The final product in all cases is a data publication, which can be standalone or in combination with a scientific article (example: http://doi.pangaea.de/10.1594/PANGAEA.762888). Data sets with insufficient metadata are rejected in general. This does not apply for selected legacy data, which, in case metadata are missing, are supplemented with a corresponding comment.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

IX. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Submitted data are processed to tab delimited text files and archived using the PANGAEA editorial system (AIP). Each data set is assigned a DOI and is registered at DataCite (http://www.datacite.org/). Except binary objects all data and metadata are imported into a relational database, which keeps the structure and semantics for each data entity. The submission and archiving workflow is documented in the PANGAEA internal Wiki. A description of the archiving environment and the corresponding workflow is given in a recent publication: https://doi.org/10.1016/j.jbiotec.2017.07.016 - Figure 6.
Data are backuped regularly on hard drives (daily incremental) and tapes (monthly). The maximum loss in case of failure of systems is one day.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

X. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Except binary objects all data and metadata are imported into a relational database, which keeps the structure and semantics for each data entity. Thus data sets inclusive data descriptions can be compiled on the fly and usable output formats can be adapted to actual standards and needs (SIP). For binary objects (e.g. images, video, geophysical data) the PANGAEA group keeps track on long term usability, in particular on the availability of software that can be used with the data objects (version changes, backward incompatibilities, new software etc.). If necessary binary objects will be transformed to newer formats.

The computer center of the AWI takes care of the proper function of hardware and software systems including backup of data and migration of data from outdated media. For documentation of all systems Confluence (Atlassian) is used.

A transfer of custody can be managed by reducing PANGAEA to a file based repository. In this case a file based copy of all data sets including possible binary object files would be transfered to the virtual cluster of the University Bremen.

A preservation plan for PANGAEA is in preparation. Our data policy (http://store.pangaea.de/Publications/PANGAEA/pangaea-data-policy.pdf) is till effective, but will be revised in 2019. A privacy policy was added in 2018 (https://www.pangaea.de/about/privacypolicy.php). Terms of usage/service are in preparation and will be integrated into the data submission components (as part of the data submission ticket).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XI. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Except binary objects all data and metadata are imported into a relational database, which keeps the structure and semantics for each data entity. Thereby the data model of PANGAEA and the functionality of the editorial system enforce that each data set has sufficient and consistent metadata (incl. provenance, listing of measurement and observation types, and methodologies used). Using our editorial system, extent and type of metadata is adapted to the scientific field and type of the data. The metadata marshalled from the RDB complies to various metadata standards (e.g. ISO19115, DataCite, schema.org, DIF, ABCD etc). Middleware components check for the correct range of values and for outliers in the data, spatial and temporal coverage is checked, consistency of the level of scale is checked, data and metadata are checked for unvalid characters, and broken links to outside references are detected (references to related literature, other versions of the data etc.). The final check for the correctness of data and metadata (e.g. the validity of used methods) is done by PANGAEA data editors who also ensure that possible community standards are regarded. In this sense PANGAEA will only take full responsibility for the technical quality. Data providers will be responsible for the scientific quality of their data.
Data users can raise issues with archived data using the PANGAEA contact which is connected with the ticket system. In addition, data sets can be rated via social netwerks including altmetrics, e.g.
https://doi.pangaea.de/10.1594/PANGAEA.816720

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:

Accept

XII. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

Archiving of data in PANGAEA follows defined workflows, which comprise technical structures, tools, and workflows as well as the organization of data and metadata by the editorial staff affiliated with PANGAEA (https://www.pangaea.de/about/team.php). Documentation is supplied in the PANGAEA Wiki (https://wiki.pangaea.de/wiki/Main_Page - for the technical part see our recent paper: https://doi.org/10.1016/j.jbiotec.2017.07.016 - Figure 6 and 9 show diagrams for ingest, archiving, and dissemination). Wiki contents are under revision. Due to the rapid changes and extensions in the scope and functionality of PANGAEA during the last years, the documentation in part is not any longer up to date.
For users archiving conditions and procedures are described on the PANGAEA data submission page (https://www.pangaea.de/submit/). This also includes an information on the scope of data accepted by PANGAEA.
The PANGAEA editorial is collaboratively organized in an editorial board with a flat hierarchy and all data editors having their specific working domain. This mostly matches their scientific expertise. Some editors are responsible for cross-cutting topics. All issues raised by editors are recorded in the PANGAEA ticket system, and discussed and decided within the editorial board.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIII. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

PANGAEA offers a variety of ways to discover data: (1) via the PANGAEA search engine (www.pangaea.de - ElasticSearch) allowing for full text and faceted searches, (2) via Google (sitemap), (3) via numerous portals harvesting PANGAEA metadata, including the GEO data portal, INSPIRE, IODP, ICSU-WDS, PubMed Central, OpenAIRE, Scopus, DataCite, DataONE, GBIF etc.). The sitemap is also used for the new Google Data Search.
All data sets held in PANGAEA are citable and accessible through a DOI. Citations can be downloaded as RIS, BibText, and text. Persistent identifiers are currently also used for literature (DOI), authors (ORCID), projects (fundRef), awards, and samples (IGSN - e.g. https://doi.pangaea.de/10.1594/PANGAEA.839477?format=html). Data sets are cross-linked with literature via Scholix. ORCIDs can be used for authentification in PANGAEA (https://goo.gl/P6rw71). Data sets in PANGAEA are cross-linked with genomic data using accession numbers (e.g. https://doi.pangaea.de/10.1594/PANGAEA.869129). Outside links are checked regularly for 404 errors once a week.
To enhance discoverability PANGAEA makes extensive usage of terminologies (included in the data model). The marshalling routines to extract metadata from our RDB includes a component for metadata annotations enabling facets and recommendations (compare https://doi.org/10.1016/j.jbiotec.2017.07.016).

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XIV. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The data model of PANGAEA and the functionality of the editorial system enforce that each data set has sufficient and correct metadata. Extent and type of metadata is adapted to the scientific field and type of the data. Because data and metadata are stored in a relational database, data sets inclusive data descriptions can be compiled on the fly and usable output formats can be adapted to actual standards and needs. Currently, data can be downloaded in a standard format as tab delimited text; data collections can be packed in zip folders. Formats for binary data are conform to community standards. In 2018 output formats will be extended by csv and a JSON based format. Metadata can be downloaded in standard formats (e.g. ISO19139: http://ws.pangaea.de/oai/provider?verb=GetRecord&identifier=oai:pangaea.de:doi:10.1594/PANGAEA.767698&metadataPrefix=iso19139). Further standards supported are DataCite, DIF, ABCD, Darwin Core, DC, schema.org. Metadata schemas are serializations, open to future adaptions and extensions.
Semantics of data sets are harmonized using standard terminologies. Consistent structures and semantics allow for effective integration of data and compilation of data products. Data integration is also supported by the PANGAEA data warehouse which can be either be used manually as part of the PANGAEA search or programmatically using the REST API. In late 2017 PANGAEA initiated a new RDA WG aiming at harmonizing measurement & observation types (https://goo.gl/ANGY5N). This will also be useful in mapping different standards of measurement & observation types (e.g. CF variables, EBV, EOV etc.)

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

TECHNOLOGY

XV. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

The basic technical structure of PANGAEA corresponds to a three tiered client/server architecture with a number of clients and middleware components controlling the information flow and quality. On the server side a RDBMS (PostgreSQL) is used for information storage. For better performance high volume and binary data (e.g. geophysics, pictures) are stored in consistent formats on hard disk arrays and tape archives. Fast access to compiled data is ensured by a data warehouse (Infobright), mirroring the data inventory. All metadata is replicated into middleware/frontend systems (Elasticsearch) for fast access and search capabilities. All public interfaces to the information system are standards conform (W3C, ISO, OGC) and are based on web services (mostly SOAP, REST) including a map supported (Google Earth, Google Maps) search engine (Elasticsearch). Metadata are dynamically marshalled from the RDB to a PANGAEA specific metadata format, stored in Elasticsearch, and transformed (XSLT, XML to JSON transform) by the frontend to various content standards (JSON-LD according to schema.org, DataCite XML, Dublin Core XML, but also more community specific ones like ISO19115/ISO19139, DIF, Darwin Core). They are disseminated via OAI-PMH (https://ws.pangaea.de/oai/), HTTP content negotiation (based on DOI standards), or other protocols. The PANGAEA ticket system used for data submissions, general user requests, or bug reports is JIRA Core (Atlassian). The PANGAEA editorial system is based on the 4th Dimension software (frontend to PostgreSQL).

All basic hard- and software services are supplied by the computer center of the AWI. Most backend/middleware systems and all front end web servers / search engine are running on virtual machines (VMware), supplying sufficient capacity and performance, as well as high availability due to virtualization. The Postgres backend database is running on a dedicated machine supplied with a high performance IO system. Machines are operated with Linux systems (Ubuntu, CentOS) and Windows Server. The editorial system is running as a client/server system on a Windows Server 2016 (VMware). Hardware is generally renewed every 3-4 years, which is transparent to the system because of virtualization. Operating systems are regularly updated to the latest releases and patches. The central archival storage system replicates up to 5 PB of data into two robotic tape archives (Sun/Storagetek SL8500), which are located in separate buildings. The backup facilities work on high capacity LTO-tape drives and media installed in this SL8500/SF15k environment. The AWI has support contracts for all relevant hard- and non-open-source software products with 5 days business hours response, part replacement 1 day. Proper functioning of PANGAEA related systems, applications, services, and processes is also ensured by usage of a professional monitoring software for IT infrastructure.

General descriptions of used systems and software are supplied within our Wiki (MediaWiki, https://wiki.pangaea.de/wiki/Main_Page), which is currently under revision. A second Wiki (Confluence) is operated separately from the other PANGAEA infrastructure for disaster safety and recovery containing internal information about servers, installation, maintenance and relationships. The technical Wiki also covers instructions for running own services, databases, data warehouse as well as topics like VM snapshots and backup.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

XVI. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository

Response:

PANGAEA related infrastructure is mirrored in two different buildings. All data is replicated between two different buildings. Regular backups on tape are deposited in a fire and waterproof safe. There are alert systems on power, cooling, and water with 24/7 stand-by service by local staff. Fire alerts are directly sent to the fire department (SLA - response < 15 min). Backend and middleware systems are protected behind a firewall, frontend systems (search frontend, web services etc.) can be accessed from outside with less limitations, but still protected by a firewall (DMZ). Front end systems do not have write access and only limited read access to the backend database and tape archives.

Backup power supply: The computer center of the AWI maintains several UPS, allowing to operate all WRMC hardware for up to 30 minutes. If the problem cannot be fixed in this time a diesel fueled emergency power generator can be started which bridges another two days.

Security of the technical infrastructure is further ensured by (1) the professional architecture and design of soft- and hardware systems, (2) short-term security patches for soft- and hardware (specific channels), (3) monitoring tools for hardware, firewall, software, services, performance, and attacks, and (4) regular trainings for technical staffs, e.g. web hacking.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

APPLICANT FEEDBACK

Comments/feedback

These requirements are not seen as final, and we value your input to improve the core certification procedure. To this end, please leave any

comments you wish to make on both the quality of the Catalogue and its relevance to your organization, as well as any other related thoughts.

Response:

R0: not sure what is meant by "Publication repository" - for literature?
R0: not clear, why the curation level is not part of R11/R12/R14 - between criteria redundancies (also between other requirements - a bit confusing)
All in all a comprehensive catalogue!

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments: