1st NFDI Conference:

Extended Abstract

Particle, Astroparticle, Hadron and Nuclear Physics accelerates the NFDI

PAHN-PaN

1. Formal details

Planned title of the consortium

Particle, Astroparticle, Hadron and Nuclear Physics accelerates the NFDI

Acronym of the planned consortium

PAHN-PaN

Lead institution or facility

Deutsches Elektronen-Synchrotron (DESY), Notkestr. 85, 22607 Hamburg
GSI Helmholtz-Zentrum für Schwerionenforschung GmbH, Planckstr. 1, 64291 Darmstadt

Name and work address of a contact person (including email address and institutional affiliation)

Dr. Thomas Schörner-Sadenius, DESY, Notkestr. 85, 22607 Hamburg
thomas.schoerner@desy.de
Dr. Thorsten Kollegger, GSI, Planckstr. 1, 64291 Darmstadt
t.kollegger@gsi.de

Members of the planned consortium (including institutional affiliation, without address)

Prof. Dr. Joachim Mnich, Prof. Dr. Marek Kowalski, Dr. Thomas Schörner-Sadenius, Dr. Christoph Wissing, DESY
Prof. Dr. Matthias F. M. Lutz, Dr. Thorsten Kollegger, Dr. Kilian Schwarz, Prof. Dr. Joachim Stroth, GSI
Prof. Dr. Karl-Heinz Kampert, Prof. Dr. Christian Zeitnitz, Universität Wuppertal
Prof. Dr. Florian Bernlochner, Prof. Dr. Günter Quast, Dr. Andreas Haungs, Dr. Andreas Heiss, KIT
Prof. Dr. Markus Schumacher, Prof. Dr. Marc Schumann, Universität Freiburg
Prof. Dr. Johannes Haller, Dr. Gregor Kasieczka, Prof. Dr. Peter Schleper, Dr. Hartmut Stadie, Universität Hamburg
Prof. Dr. Andre Brinkmann, Prof. Dr. Volker Büscher, Prof. Dr. Frank Maas, Prof. Dr. Hartmut Wittig, Johannes Gutenberg-Universität Mainz
Prof. Dr. Thomas Kuhr, Dr. Günter Duckeck, LMU München
Prof. Dr. Ulrich Uwer, Prof. Dr. Silvia Masciocchi, Prof. Dr. Hans-Christian Schultz-Coulon, Prof. Dr. Stefan Wagner, Universität Heidelberg
Prof. Dr. Martin Erdmann, Prof. Dr. Alexander Schmidt, Prof. Dr. Achim Stahl, RWTH Aachen
Prof. Dr. Arnulf Quadt, Universität Göttingen
Prof. Dr. Siegfried Bethke, Dr. Stefan Kluth, Max-Planck-Institut für Physik München
Dr. Elena Bratkovskaya, Prof. Dr. Hannah Elfner, Prof. Dr. Carsten Greiner, Prof. Dr. Owe Philipsen, Prof. Dr. Dirk Rischke, Prof. Dr. Marc Wagner, Johann Wolfgang Goethe-Universität Frankfurt
Prof. Dr. Hans-Werner Hammer, Prof. Dr. Dr. h.c. Norbert Pietralla, Prof. Dr. Achim Schwenck, Technische Universität Darmstadt
Prof. Dr. Frithjof Karsch, Dr. Olaf Kaczmarek, Dr. Christian Schmidt, Universität Bielefeld
Prof. Dr. Reinhard Beck, Prof. Dr. Klaus Desch, Prof. Dr. Dr. h.c. Ulf G. Meißner, Prof. Dr. Carsten Urbach, Universität Bonn
Prof. Tom Luu Ph.D., Forschungszentrum Jülich
Prof. Dr. Nora Brambilla, Torsten Dahms Ph.D., Prof. Dr. Laura Fabbietti, Prof. Dr. Stephan Paul, Prof. Dr. Antonio Vairo, Technische Universität München
Prof. Dr. Uli Katz, Friedrich-Alexander-Universität, Erlangen-Nürnberg
Prof. Dr. Peter Buchholz, Universität Siegen
Prof. Dr. Thomas Trefzger, Universität Würzburg
Prof. Dr. Kevin Kröninger, Technische Universität Dortmund
Prof. Dr. Dominik Stöckinger, Technische Universität Dresden

Prof. Dr. Lorenz von Smekal, Prof. Dr. Claudia Höhne, Justus-Liebig-Universität Gießen
Dr. Andre Sternbeck, Friedrich-Schiller-Universität Jena
Prof. Dr. Thomas Stöhlker, Helmholtz-Institut Jena
Prof. Dr. Ulrich Wiedner, Ruhr-Universität Bochum
Prof. Dr. Gunnar Bali, Prof, Dr. Andreas Schäfer, Prof. Dr. Christoph Lehner, Universität Regensburg
Prof. Dr. Andrey Surzhykov, Technische Universität Braunschweig und Physikalisch-Technische Bundesanstalt
Prof. Dr. Burkhard Kämpfer, Helmholtz-Zentrum Dresden-Rossendorf
Prof. Dr. Anton Andronic, Prof. Dr. Christian Klein-Boesing, Universität Münster
Prof. Dr. Michael Schmelling, Max-Planck-Institut für Kernphysik Heidelberg
Prof. Dr. Andreas Zilges, Universität Köln

Participants in the NFDI conference (names, institutional affiliation and email address; max. 3 persons)

Dr.Christoph Wissing (DESY, *christoph.wissing@desy.de*), Prof. Dr. Thomas Kuhr (LMU München, thomas.kuhr@lmu.de), Dr. Kilian Schwarz (GSI, *k.schwarz@gsi.de*)

Research area of the planned consortium (research area according to the DFG classification system [not subject areas]:

32 (Physics)

Participating research institutions (without address)

CERN (contact person: Dr. Markus Elsing, markus.elsing@cern.ch)

Participating infrastructure facilities and/or potential information service providers (without address)

DESY, GSI, KIT, Johannes Gutenberg-Universität Mainz

Planned proposal submission date (2019, 2020, 2021)

2019

Overview diagram or organisational chart for the planned consortium

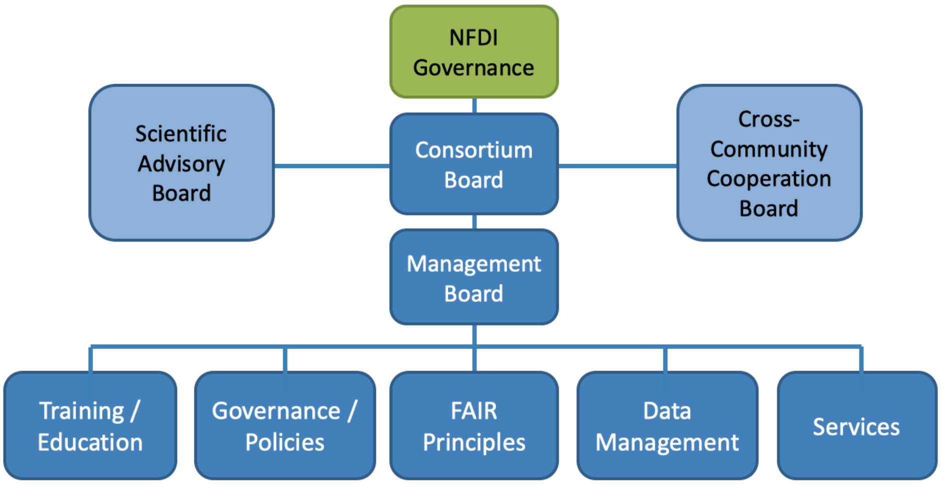

Figure 1: The foreseen governance structure of the PAHN-PaN consortium.

The consortium plans to act in five work packages (WPs). It will be governed through the consortium board, while the day-to-day project management is handled by the management board. In the management board, the WP leaders and two directors of the project are present.

2. **Subject-specific and infrastructural focus of the planned consortium**

- Key questions/objectives of the consortium

The mission of the communities represented in this consortium is to understand the fundamental laws of nature at the microscopic level and to apply them to understand the origin and evolution of the universe. These communities always have been at the forefront of technological developments – for instance, these days the 30th anniversary of the world-wide web WWW, which was invented at CERN, was celebrated. However, the development of new accelerators, new observatories and new detectors with much higher resolution and trigger rates leads to a constant drastic increase of data and subsequently to ever increasing demands on data analysis power and data management capabilities. This tremendous increase of data – we are talking about more than 1000 PByte (1 ExaByte) of scientific data per year at present and upcoming facilities – requires many new ideas, concepts and methods in order to fully harvest the science. New approaches to problems in scientific computing are needed, which again makes the involved fields drivers in IT technology. This also applies to other communities with large data sets like e.g. astronomy. This consortium represents, through the Komitees für Elementarteilchenphysik (KET), für Hadron- und Kernphysik (KHuK) and für Astroteilchenphsyik (KAT), about 3300 scientists (doctoral degree) in Germany from universities, the Helmholtz Association, and the Max Planck Society – institutions that have been collaborating since many years. The research instruments are experiments at large-scale research infrastructures like the Large Hadron Collider, experiments at CERN's Super Proton Synchrotron, at KEK's SuperKEKB accelerator, or at the upcoming FAIR[1] accelerator at the GSI, the observatories Pierre Auger, CTA, IceCube, KM3NeT, the KATRIN experiment, or the infrastructures for present and future neutrino physics and dark matter searches. Also university-based accelerators like MAMI or , in future, MESA in Mainz and S-DALINAC in Dresden are involved.

This consortium, therefore, has long-lasting experience in the analysis and the overall management of large quantities of research data: Its particle physics partners have significantly contributed to the very successful running, for more than 10 years, of the distributed computing environment for the LHC data analysis, called the "worldwide computing grid for the LHC" (WLCG) with more than 160 sites involved worldwide. Nevertheless, the upcoming experiments at the High-Luminosity LHC (HL-LHC) or at the FAIR facility and the large-scale observatories will deliver an amount of scientific data that requires fundamentally new concepts and strategies to manage, access, analyse and publish data and the results obtained from them while keeping costs as low as possible.
In astroparticle physics, the idea to combine measurements of several observatories – the so-called multi-messenger approach – will, among other things, require significantly more data, in addition to better-adapted methodical approaches.

In conclusion, it is highly important to create synergies through increased interdisciplinary cooperation. This encompasses the following steps:

- development of interdisciplinary cloud-enabled methods for big data analysis (analysis pipelines) based, among others, on "machine learning", and visualization tools;
- making the data "*fair*", which means providing, for instance, well-suited data structures and meta-data systems as well as methods for open data and open access;
- making not only data, but also analysis and simulation software packages commonly available through public catalogues;
- development of data management solutions from data ingest to data retrieval and archive, in particular for long-term data and analysis software preservation;
- development of interdisciplinary education and training methods for the digital scientist;
- development of policies and workflows for the data lifecycle process including interdisciplinary systems for generalised publications of scientific data (DOI, Orcids, etc.)

[1] In this context, "FAIR" refers to the "Facility for Antiproton and Ion Research", the accelerator complex under construction at the GSI. It has to be distinguished from "*fair*", standing for "*findable, accessible, interoperable and reusable*", see footnote (2).

- Known needs/current status of research data management in the relevant discipline/subject-specific relevance of the planned consortium:
    - From a research perspective
    - In terms of available information providers and services

Research perspective:
The mission of the physics pursued in this consortium is to understand the fundamental laws of nature at the microscopic level and to apply them to understand the origin and evolution of our universe.

In particle physics, the so-called standard model (SM) has over many decades evolved through both the development of new theoretical concepts and a large variety of experimental measurements; many aspects of this model have been confirmed with high precision, and many predictions were later spectacularly confirmed by experiments. However, there are strong reasons to believe that the SM is just an "effective theory" that produces reliable results only at the energy and length scales that we currently probe in our experiments. It seems as if the SM is embedded in a more comprehensive and richer theory that provides solutions to some of the critical problems that pose some of the main challenges for particle physics:

- The SM does not contain a candidate particle for the dark matter observed in the universe.
- We are also lacking an understanding of the origin of mass and of electroweak symmetry breaking, and of the structure of the vacuum.
- We do not have an explanation for dark energy.
- We also cannot explain the asymmetry between the amounts of matter and antimatter observed in the universe.

The mission of hadron and nuclear physics is to unravel the nature of strongly interacting matter. It is important to study how complex structures are formed in terms of the fundamental constituents – quarks and gluons. The relevant and effective degrees of freedom need to be identified in order to describe and predict empirical cross sections as measured in the laboratory. Extreme states of matter, which have analogies in cosmic evolution, are formed in the collision zone of colliding nuclei if the total energy in the center of mass substantially exceeds the combined mass of the collision partners. The properties of these transient states are revealed by investigating the particles emerging from the collision zone with the help of complex detector systems. Important questions addressed concern the formation of matter during hadronization, the role of confinement and spontaneous chiral symmetry breaking therein, the nature of the Quark Gluon Plasma, and the possible existence of so far unknown phases of matter. Further research interests are to address a large range of essential questions concerning nuclear structure and dynamics, nuclear astrophysics, tests of fundamental interactions and symmetries.

Astroparticle physics combines our knowledge of the largest structures in the universe with that of the smallest building blocks of matter and the forces between them. It is a fascinating field of science living at the interfaces of astronomy, astrophysics, cosmology, elementary particle physics and nuclear physics. Activities in this young research field have increased dramatically in the last two decades, with an emerging challenge to combine measurements of several observatories. This multi-messenger ansatz is expected to lead to a better understanding of high-energy processes in the universe. Currently this is hampered by low statistics (an obstacle that will be overcome by the next generation of observatories), by uncertainties in the reconstruction of these rare events, and by limitations in data analysis.

Information and service perspective:
Access to an efficient and powerful “information and communication technology” (ICT) infrastructure is an elementary precondition for a sustainable success of the participating

disciplines. Therefore, the installation of a national research data infrastructure (NFDI) is very important for the communities involved in the consortium.

Due to the large complexity and costs and – most importantly – lifetimes of our experiments, it has become one of the major characteristics of our fields of research that large international collaborations consisting of many laboratories and universities are formed. Therefore, it is in our consortium’s vital interest to establish and maintain strong partnerships worldwide, which is sometimes a challenge in itself, and to continue playing a leading role internationally. This will remain true for the next projects to come that might represent even larger endeavours. Germany has committed itself to long-term responsibilities in many international partnerships and collaborations in physics: at the LHC, at Belle II, in CTA, at FAIR and elsewhere. These responsibilities are fixed in collaboration agreements and acknowledged and supported by the funding agencies. Therefore, the research programmes are scheduled for long time-lines and are, in large parts, decided for 20 years to come.

With the tremendously growing resolutions of the detectors at our facilities, the strongly increasing digitisation, and emerging new possibilities for simulation and data analysis offered by the development of computer technologies, there is a significantly growing demand for a large-scale federated computing infrastructure, which in turn requires a elaborate research data management. The central question resulting from this is the requirement for a reasonably priced, sustainable and efficient compute infrastructure that satisfies the demand and that can help to develop future-oriented concepts. It is particularly important to further develop the corresponding experiment software, and also the development of new and innovative software will be necessary. Development of data reduction, data compression and data filtering methods to reduce the data volumes is equally important. The information infrastructure in Germany within this research field is already outstanding. It is provided by the Helmholtz centres DESY, GSI, KIT, by Max Planck institutes, and by various universities. However, for the future it is necessary to provide an efficient integration of so-called opportunistic resources (e.g. scientific and commercial cloud providers, larger HPC centres as well as models like volunteer computing). One more important point is the concentration to few large centres storing and providing experimental data that together can be seen as a single virtual data centre. Such a centre needs to be equally accessible via all state-of-the-art access methods. For this to be possible, it is required that the compute centres for all experiments are interconnected with a bandwidth of 100 GB/s or more, at the latest from 2020 onwards.

This consortium specified the demands and developed ideas which should be dealt within the context of NFDI – see below. These ideas are in line with the „Digitale Agenda“ of the German federal government. They specify which technical and organisational measures should be taken – in close collaboration with the universities and the research centres – and which measures are of trans-disciplinary interest in the context of a future-oriented digitisation of German science and society.

The general overarching idea of all suggestions is a strengthening of science. The further development of infrastructures will also further strengthen Germany’s position in the international scientific landscape. Additionally, the participating communities of the consortium already today provide important contributions to the development of technology and to the education of qualified personnel – an aspect of decisive importance for Germany as a competitive industry location. The participation of the public in elementary research will strengthen the acceptance within the society and will also serve the promotion of young academics.

With respect to the view of information providers, the long lead times and lifetimes of large-scale projects require further measures on sustainability, long-term reliability and, last but not least, long-term funding. Note that the design, funding and construction of the LHC took about 25 years, and data taking will also last at least 25 years!

- Summary of the planned research data infrastructure that is specifically intended to address the needs of research users in their respective work processes

  Federated data management systems across disciplines and provision of data in a standardized way following the “*fair:* (*findable, accessible, interoperable, reusable*) data principles are not or not yet sufficiently available and still require many developments. In general, a modular setup should be applied so that community-specific building blocks can be replaced on demand by more fitting bricks. Wherever possible, such a system should be based upon existing tools and standards rather than on the development of new tools.
  Examples of necessary developments and key elements of a research data infrastructure proposed by this consortium are:
  - development of federated infrastructures (data lakes, hybrid models) based on cloud technology and conceptionally open for a wide range of communities and with industry;
  - development of methods for including opportunistic resources (scientific and commercial cloud systems, HPC centres, ...);
  - improvement and adaption of new methods and algorithms to the analysis and data management software, including e.g. methods of machine learning, visualisation, etc.;
  - development of intelligent methods for data reduction, data compression and data filtering to reduce the amount of data;
  - development of cross-community meta-data systems;
  - development of user-friendly catalogues and portals for data repositories and for software;
  - creation of (virtual) competence centres for „scientific data analytics“;
  - development of sustainable principles for data and software preservation and archiving according to the “*fair*” principles;
  - development and adaption of a common publication process of data and scientific results.

- Description
  - of data types
  - of underlying data processing / data analysis methodologies

  The huge amount of scientific data coming from the experiments or the simulations are well structured and annotated by meta-data. Besides experiment-specific data formats, HDF5 plays a major role. Data structures also need to allow parallel data streams. The research data need to be documented and archived in a sustainable way, which also includes analysis software preservation. In order to be able to share the research data, standards need to be defined for data formats and meta-data. Moreover, data need to be published following the “*fair*” principles. Policies and workflows for the data lifecycle process need to be developed, especially also for the generalised publications of scientific data. Close collaboration with existing data life-cycle labs in the Helmholtz Association are planned.

- Planned implementation of the “*fair*” principles[2] and information about any existing policies or guidelines in the relevant discipline

  Research data form the basis for the scientific harvest of our research. Therefore, the analysis, the documentation and the safe archiving of these data are essential, including archivable virtual research environments. Data are often shared within collaborations in order to create synergy and new insights through collaborative work. In order to allow all participating scientists access to these data, it is necessary to agree on standards for data formats and also on the format of meta-data. The effectiveness of such measures will show as soon as scientists try to access earlier analysis results. It is increasingly important to publish research data following the aforementioned “*fair”* data principles. This increases the visibility of research activities and enables third-party scientists to reuse any data set. It may, however, also require well-adapted models for licensing and also the possibility to cite data sets. All these aspects have to be considered within an overarching concept for

[2] https://www.force11.org/fairprinciples

research data management. In order to reduce the corresponding effort, it is necessary to digitize all processes from data taking up to publication.
As a result, key elements of a "*fair*" infrastructure proposed by this consortium are:
- development of cross-community meta-data systems;
- development of user-friendly catalogues and portals for data repositories and for software;
- development of sustainable principles for data and software preservation and archiving according to the "*fair*" principles;
- development of workflows for open data, open access, DOIs, publications etc.

- Planned measures for user participation and involvement

In the communities represented by this consortium, there are many organisational measures already in place to assure user participation and involvement on a national and international level. As one out of many examples, the three committees KET, KAT, and KHuK have to be mentioned. The representatives are elected by the members of the respective communities on a regular basis. A step to get more users acquainted with big data analytics would be the creation of (virtual) competence centres for "scientific data analysis".
The education and training of young academics is very important for the sustainability of implemented measures and also for Germany as an industry location. An NFDI can contribute in an essential way through
- providing access to resources;
- providing methods for interdisciplinary work;
- development of coordinated education and training concepts for usage of modern IT technology by young academics;
- development of coordinated curricula for the "data scientist" which would stress, e.g. the handling of experiment data and methods for data analysis;
- better reputation and career possibilities for ICT-close personnel in individual subjects;
- support of graduate schools, summer schools, MOOCs, hackathons for scientific computing and data mining;
- sustainable user support and user training;
- providing additional sustainable resources for user support;
- funding of educational outreach programmes for the interested public.

- Existing and intended degree of networking of the planned consortium
  - nationally (in particular with other, potential future consortium or existing state-level initiatives)
  - internationally
  - between the infrastructure facilities and the research community
  - with respect to major networking topics

Already today, the consortium is very well connected to other research fields on a national and international level, in particular with astronomy. Large observatories like the upcoming Square Kilometer Array (SKA) or the Cherekov Telescope Array (CTA) are suffering from the same problems as facilities in particle physics or in hadron and nuclear physics, and there are already many cooperations in place through EU-funded projects like ESCAPE[3] or "extreme data cloud" (XDC), or directly through CERN. There are also strong links to the photon and neutron communities already now. Sites like DESY, GSI or KIT as big labs and as infrastructure and facility providers are supporting many communities beyond this consortium and act as multipliers. For all communities in this consortium, it is essential to be embedded in the national (e.g. ErUM-Data) and the international context (e.g. EOSC). Members of this consortium are active in many national and international activities like EOSC or the GO-FAIR initiative.
Moreover, the large facilities of the participating communities are already embedded in international collaborations. This is true especially for the experiments at the LHC at CERN,

[3] https://escape2020.eu

for Belle II in Japan, for observatories like CTA, or for the experiments at the FAIR accelerator in Darmstadt. For this consortium, a national research data management (NFDI) in Germany has to consider these international aspects from the very beginning in order to be successful. Furthermore, a national governance structure for the NFDI has to be created that includes the communities and that has to be able to react in a flexible way to developing needs from the user side. Here it is necessary to find community-specific as well as common solutions. Examples for these aspects are:

- development of data policies, e.g. access rights, embargo periods etc.;
- copyrights for data and software need to be developed; one needs to be able to distinguish between data and the conceptual achievements of individual scientists;
- setup of a management infrastructure for using scientific services;
- development of concepts for open-access publishing;
- integration into and active participation in a European science cloud strategy;
- promotion of national and international collaboration.

For the consortium, cooperation with other disciplines of research is an integral part of science. Especially the cooperation with mathematics and with computer science could lead to win-win situations for reasonable scientific data analysis and in the development of methods and algorithms that deal successfully with the upcoming enormous amounts of data. Furthermore, it is required that domain-specific research is closely linked with the work of developers and operators of large facilities. Examples of such collaborations are:

- joint interdisciplinary development of new algorithms and tools for, e.g., “image processing” among physics, mathematics and computer science;
- introduction and development of parallel algorithms and methods for machine learning, data mining, real-time simulation and especially for the analysis of huge amounts of data;
- close connection to the national and international large research facilities and industry.

Figure 2 shows the overall embedding of the consortium into the scientific environment.

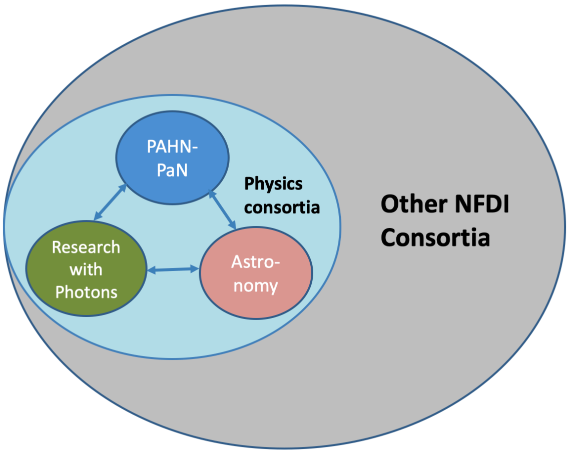

Figure 2: The embedding of the consortium (blue ellipse) in the nearer environment of other physics-related consortia and the wider scientific context.

- Additional information

The participating communities represented through the committees KHuK, KET, and KAT formed a consortium that intends to create and further develop an interdisciplinary, community-overarching research data management infrastructure adjusted to the developing needs of the connected fields of science. The planned research data infrastructure will be based on the experience obtained with the already existing infrastructures and computing environments of all participating communities. All participating research institutions have long-standing experience in the production, distribution and analysis of large amounts of data. Through close cooperation among all involved partners, this experience shall be made available for the benefit of the NFDI and other communities, as well as transferable to industry and society. More interested participants of the participating communities are invited to join and actively contribute to this consortium.