# Proposal for ICSU-WDS/RDA WG on a Framework and Registry for Brokering and Mediation Components

# Background

## Mediation and Brokering

In a SOA framework, software components interoperability is implemented by defining and using common protocols. In a web services framework, interoperability protocols are characterized by their interface methods, and bindings as well as their payload content. For example, in the case of protocols for data discovery and access, the payload content contains data and metadata encoded using specific models.

Mediation and adaptation modules are often used to map two different content models or two different interface methods or two different binding types. Commonly, a mediation module addresses the mediation of one feature characterizing two different protocols -e.g. the payload content model.

Brokering services can be used to implement more advanced and general mediation functionalities. Brokering components address all the three protocol heterogeneities: methods, bindings, and payload content models. In addition, they implement mediation from many-to-many different protocols

In the geospatial information domain, Brokering services were successfully introduced to mediate across the different disciplines/Community protocols for data discovery, access, transformation and (in the most advanced cases) processing.

## Problem Description

The scope of the WG covers data sources from research and scholarly communication. Within this domain obstacles for wider application of brokering techniques are:

1. Multiple service protocols for data discovery, access, and application or processing;
2. Multiple content standards for data and metadata, augmented by Community profiles and non-standard implementations;
3. Multiple vocabularies and ontologies.
4. Multiple adaptation and mediation modules that are not guaranteed to be compatible.

Research projects and research data infrastructure initiatives often solve problems associated with this diversity as a matter of course, but the knowledge gained and components developed during such a process are not visible and useful to others. Furthermore, project life cycle limitations lead to lack of sustainability, loss of expertise, code, and infrastructure.

## Address the Need

* Define a description schema for services, vocabularies, ontologies, content standards, and adaptation components based on current community research and operational developments that allow services and clients to be more effectively interfaced with mediation components; leverage prototype registries developed through RDA and other activities based on the above.
* Describe a collection of existing mediation and adaptation components that can interoperate through well-defined existing interface specifications and applicable standards.

* Create a mediation and adaptation components registry -the objective is to support implementation of a more general and agnostic mediation capability. This registry outcome may be integrated into DTR and other registry activities in RDA once its characteristics are defined and tested.
* Define a test bed environment for testing interoperability of mediation alternatives leading to recommendations for application areas. The focus will be on metadata and data mediation across data systems that address different disciplines and scopes.

Mediation functions to be assessed by the test bed:

1. Data discovery and access protocols including harvesting and synchronous distribution (subscription/notification).
2. Data and metadata content transformation (harmonisation).
3. Metadata content enhancement and Linked Open Data enablement through vocabularies and ontologies.
4. Application to popular protocols and service definitions.

## Support of RDA Working Group Outcomes

|  |  |
| --- | --- |
| **RDA Working Group Outcome** | **Brokering Framework WG Alignment** |
| New data standards or harmonization of existing standards. | New standard(s) will likely be needed to describe a component within the typical research data infrastructure in a standardised manner. This ensures that its capabilities and role within the architecture is clear and unambiguous. |
| Increased data sharing, exchange, interoperability, usability and re-usability. | Interoperability automation is strongly dependent on the availability of brokering and mediation components, and on its automated discovery and invocation. |
| Increased discoverability of research data sets. | Brokering exposes data catalogues that are not completely standardised and thereby increases the scope of discovery. |
| Better management, stewardship, and preservation of research data. | No direct contribution. |

# WG Charter

The workgroup has the following vision and mission:

## Vision

*“To define, develop, test, and implement a Brokering Framework that allows publication, discovery, and invocation of mediation and adaptation components in a standardised manner.”*

## Mission

1. Community consensus on the way in which mediation and brokering components are described and invoked, based on real use cases – “Brokering Framework” – leading to a formal way for description and invocation.. The formalisation of a standard in this area is an external process and may involve participants in the working group, but is not a task for the working group;
2. Demonstrating the value of the framework by developing, testing, and commissioning a reference implementation of a mediation registry.

## Membership

Membership is open to any participant in RDA, and will include at least the initial membership (see below). Broad experience and geographic participation will be encouraged. Organizations such as WDS with broad geographic and organizational representation will be invited to participate.

## Methods

The working group will make use of the following methods:

1. Contributions from members in respect of current working components and use cases, through regular working group meetings. Such meetings will be held every 2 months and, where possible, aligned with RDA Plenary Meetings.
2. In-kind development contributions. Provisionally, this will include work to be done by CNR and DIRISA. Other in-kind contributions can come from GEOSS, PANGAEA and possibly OpenAire.
3. Leveraging and collaborating in developments and outcomes of RDA working groups in areas of registries, metadata, data fabric and others.

# Value Proposition

## A Common Architecture [3]

The generalized scientific data use case is usually implemented as one or more components of a ***Research Data Infrastructure (RDI)***, and does so within a particular architectural pattern. This is so because many of the existing protocols, standards, and specifications for a Global Research Data Infrastructure have been developed with a specific architecture in mind.

This pattern can generally be described as a service-oriented ‘Publish, Find, and Bind’ architecture [4]. In such an architecture, owners and creators of data and data services have the obligation to publish it, but the obligation is on the end user or system to find the data that is required, and to ‘bind’ it to a particular end use – whether this is visual inspection, download, online processing, or analytics.



Within the ‘Publish, Find, and Bind’ architecture, there are a number of components and resources that are critical, and some that are optional. It is critical, for example, that ***published*** data exists, is described in such a way that it can be ***discovered***, and that some form of ***access*** is provided – even if the access is manual and by way of personal request. And, it is non-negotiable that the data, once applied, be ***cited*** properly.

Other aspects, such as proper ***curation***, and the ability to automate tasks such as ***visualization***, ***exploration***, ***analysis***, and ***processing***, are optional.

Lastly, there are two kinds of ***brokering*** to consider in the generalised architecture. Implementation success in this architecture is strongly dependent on ***standardisation of interfaces*** for publication, finding, and binding. In practice, complete standardisation is unlikely to ever be achieved particularly across diverse disciplines, and hence brokering is required to modify variations of a standard or non-standard data, metadata, and application elements. In general, ***brokering*** describes adjustments in the service syntax, while adjustment to semantics and schema or content is called ***mediation***. These are both considered in the development and outcomes of this working group.

It is possible to conceive that with assistance of brokering, end users will have a choice of more or less equivalent data services, discovery channels, and functionally equivalent applications. In such a scenario, the ability to assess the resource for usability and quality, metrics in respect of cost (if any) and service duration or latency could become increasingly important. A second type of brokering, based on a variety of metrics and aimed at ***optimization*** of online, automated processing, is a logical future use case.

Hence, from a theoretical perspective, it is evident that brokering and mediation have an important role to play in this commonly adopted architecture.

## Evidence of Need

This need was identified in the GRDI 2020 project, and the final report [1] confirms the significant role that brokering and mediation is expected to play in future:

*“The ultimate aim should be the definition and implementation of an* ***integrated mediation framework*** *capable of providing the means to handle and resolve all kinds of heterogeneities and inconsistencies that might hamper the effective usage of the resources of a global scientific data infrastructure information infrastructure [2]. We envision that one of the most important features of the future scientific data infrastructures will be the mediation software.”*

and

*“We, thus, envisage that one of the most important features of future disciplinary data infrastructures will be the efficient implementation of a set of boundary objects defined by the members of the disciplines being supported.”*

In this context, the ‘boundary objects’ are equivalent to brokering and mediation components [5].

This last statement touches on an important aspect to consider: that any brokering and mediation framework should include the means for the community to easily share their brokering and mediation expertise – this is the only way in which a scalable solution can be achieved.

## Practical Outcomes

The outputs from the Brokering Framework Working Group should include:

1. Confirmation of the elements of a brokering framework: component description standard, registry specifications
2. A community consensus, achieved via RDA working group efforts, to develop and publish a brokering and mediation component description standard that can be used as the basis of a registry of such components.
3. Community consensus on the capabilities (service methods) of a registry: allowing discovery and description of a brokering and mediation component. This work is best done in collaboration with other RDA working groups, leveraging existing initial contacts through the Broker Interest Group over the last year.
4. A shortlist of important components that are currently operational will be obtained with community assistance, and populated in the registry.
5. Testing the registry in the context of real-life applications: DIRISA[[1]](#footnote-1) will be implementing a registry that can be duplicated as part of a test bed for the Brokering Framework Working Group.

# Existing Work

ESSI Lab (CNR) and collaborators [6] have done considerable work to develop brokering approaches for GEO (GEOSS Broker) within the EuroGEOSS project. The value of the EuroGEOSS brokering approach has been demonstrated in extending the data resources available through the Global Earth Observation System of Systems (GEOSS) from a few hundred to over 28 million in a matter of 3 months. Brokering offers therefore a real chance to facilitate truly multi-disciplinary big data science and address the scientific challenges of our time. This work has since been formalized in the GI-CAT [7] suite of applications, and the Brokering Framework Working Group will aim to incorporate the brokering components into the registry. GI-CAT is currently in use by GEO, EU-BON, GMOS, ACADIS, PANGAEA, SeaDataNet, and others. Specific work has also been done to map CERIF to INSPIRE that can serve as an example of detailed brokering and mediation [14].

EarthCube [8] has developed the BCUBE component for brokering, based on GI-CAT, and focuses on the domains of Hydrology, Oceanography, Cryosphere and Climate. As such, EarthCube collaboration will contribute detailed brokering examples and use cases.

Within the EU, there are two initiatives that currently provide brokering services to a wider research community: OpenAire [9] and EUDAT [10].

OpenAIRE has built up a software framework collecting metainformation from from numerous content providers. Guidelines supplied for content providers are widely adopted and include recommendations on export models (formats and semantics) for scholarly communication data sources. The framework comprises:

* D-NET Software Toolkit (**useful for development of WG prototypes**): production-ready technology that implements the “brokering and mediation” patterns (registry of resources, discovery of resources, workflow creation, scheduling and automated execution, and customization of resource types)
  + http://www.d-net.research-infrastructures.eu

DLI-Service

* Models and format for exchanging data-publication links
* Prototype (soon production) service:
  + http://dliservice.research-infrastructures.eu

EUDAT has elaborated a number of infrastructural tools among them a metadata discovery service - B2Find [18] - which is used to harvest metadata from research data collections from EUDAT data centres and other repositories. EUDAT is open to collaboration and will be approached to participate formally in the working group.

DataOne [11] includes support for ILTER, and as such brokers a variety of global data providers already. DataOne is a collaborator in the EUBON project [16], and as such, either of these two initiatives will be able to contribute a significant body of knowledge to the brokering registry.

PANGAEA has set up a brokering framework – PanFMP [19] - applicable to earth and environmental sciences [17]. The framework is used since 2007 for the ICSU World Data System (WDS) data portal [20] and for various community portals, e.g. more recently for the German Federation for Biological Data (GFBio) data portal[21]. The framework can be combined with other frameworks like the GI-CAT or GeoNetwork [22].

SAEON [12] has developed brokering capabilities for a number of popular protocols and meta-data schema and map these onto extended DataCite meta-data schema for internal indexing and DOI registration. These components will be made available via the new DIRISA project.

While the existing work described above focuses on brokering developments, there have been a number of RDA outcomes in the areas of registries, metadata, etc that will play a role in the creation of brokering component registries. There are also application areas such as wheat interoperability that have discipline focused solutions which will be invited to participate in populating the test bed environment. Use cases in programs such as EarthCube and ENVRI plus will be examined for appropriateness in the context of the test environment.

# Work Plan

1. Confirm use cases and examples of brokering mediation, with a view to classifying them and developing a data model for description of components.
2. Develop content and service standards for a registry of brokering and mediation components.
3. Create and populate a registry of mediation options that allow components to be shared and improved. Wherever possible, this could be based of contructs for existing RDA outcomes.
4. Create a test environment to examine existing and future capabilities.
5. Test and evaluate.
6. Consider governance of the registry and the test environment in collaboration with the Brokering and other RDA Interest/Working Groups.

## Deliverables and Milestones

|  |  |  |
| --- | --- | --- |
| **Task or Deliverable** | **RDA Plenary** | **Expected Timelines** |
| Use cases and classification of brokering and mediation components | 8 | 3 months |
| Develop content and service standards | 9 | 9 months |
| Create a registry of mediation options | 9 | 9 months |
| Test environment - initial/third party implementation | 10 | 12 months |
| Develop protocols for the test environment | 10 | 12 months |
| Contingency and slippage provision for the above: final delivery | 10 | 15 months |
| Governance of registry and test environment | 10 | 15 months |
| Recommendations and report | 11 | 18 months |

# Adoption Plan

Our adoption plan is summarized below.

|  |  |  |
| --- | --- | --- |
| **User Community** | **Nature of Adoption** | **Outreach Method** |
| RDI Developers | Registry service calls embedded into code, as a way of identifying brokering and mediation options and offering these to end-users. | RDA Communication  Reference Implementation |
| RDI Developers | Avoid duplication of effort by discovering components and improving/ extending rather than redeveloping | RDA Communication  Links to Open Code Repositories |
| RDI Developers |  |  |
| End Users | Describe and encode brokering and mediation work and publish as a research output with Permanent Identifier. | RDA Communication  Registry usage guidance  Communication to major infrastructure projects (EUDAT, DataOne, GEO, EarthCube, EU BON, ANDS, …) |
| End Users | Benefit from vocabulary and ontology mediation reliably embedded into meta-data capturing and discovery tools. | RDA Communication  Reference Implementation |
| End Users | Possibility of real-time discovery of visualization, analysis, and exploration tools to apply to open data. | RDA Communication  Reference Implementation |
| RDI Architects | Identify scope of gaps and overlaps in availability of brokering and mediation components | RDA Communication  Reference Implementation |

# Initial Membership

* Michael Diepenbroek (PANGAEA)
* Wim Hugo (SAEON, ICSU-WDS SC)
* Stefano Nativi (CNR/ ESSI Lab)
* Jay Pearlman (EarthCube, IEEE, J&F Enterprises)
* Paolo Manghi (OpenAIRE)
* Uwe Schindler (PANGAEA, Apache Software Foundation)

# References

[1] NCR-ISTI, “GRDI2020 Final Roadmap Report Global Research Data Infrastructures: The Big Data Challenges”, European Commission, February 2012, <http://www.grdi2020.eu/Repository/FileScaricati/e2b03611-e58f-4242-946a-5b21f17d2947.pdf>

[2] M. Stollberg, E. Cimpian, A. Mocan and D. Fensel, “A Semantic Web Mediation Architecture”, in: Proceedings of the 1st Canadian Semantic Web Working Symposium (CSWWS 2006), 22 pp., Springer, 2006; cited in [1]

[3] Hugo, W., “Mediated, Brokered Interoperability For Global Environmental Data Infrastructure”,SciDataCon 2014, New Delhi, India, 3-5 November 2014.

[4] P. Maue, and S. Schade, “Data Integration in the Geospatial Semantic Web”, in Y Kalflogou, “Cases on Semantic Interoperability for Information Systems Integration: Practices and Applications”, IGI Global, 2010.

[5] S. Star, “The structure of ill-structured solutions: Boundary Objects and heterogeneous distributed problem solving” in Readings in Distributed Artificial Intelligence, 1989, cited in [1]

[6] S. Nativi, M Craglia, and J. Pearlman, “The Brokering Approach for Multidisciplinary Interoperability: A Position Paper”, <https://rd-alliance.org/system/files/filedepot/97/281-1422-1-PB.pdf>

[7] <http://essi-lab.eu/do/view/GIcat>

[8] <http://earthcube.org/group/bcube>

[9] <https://www.openaire.eu/>

[10] EUDAT, “What is B2Find?” <http://eudat.eu/services/b2find>

[11] <https://www.dataone.org/working_groups/cyberinfrastructure>

[12] <http://saeos.dirisa.org/documentation/for-data-providers/G336.4.4.4%20Options%20for%20Meta-Data%20Exchange.pdf/view>

[13] <http://www.dirisa.ac.za/>

[14] E. Boldrini, D Luzi, S Nativi, and F Pecoraro, “Integrating CERIF Entities in a Multidisciplinary e-infrastructure for Environmental Research Data”, Procedia Computer Science, Volume 33, 2014, [doi:10.1016/j.procs.2014.06.031](http://dx.doi.org/10.1016/j.procs.2014.06.031)

[15] N. Houssos, B. Joerg, J. Dvorák, “OpenAIRE Guidelines for CRIS Managers 1.0”, OpenAire, 2015. <http://doi.org/10.5281/zenodo.17065>

[16] <http://eubon.eu/show/outcomes_2739/>

[17] Schindler, U, Diepenbroek, M, 2008. *Generic XML-based Framework for Metadata Portals.* Computers & Geosciences 34 (12), 1947-1955. [doi:10.1016/j.cageo.2008.02.023](http://dx.doi.org/10.1016/j.cageo.2008.02.023)

[18] <https://eudat.eu/services/b2find>

[19] <http://www.panfmp.org/>

[20] <https://www.icsu-wds.org/services/data-portal>

[21] <http://www.gfbio.org/data-portal>

[22] <http://geonetwork-opensource.org/>

1. Data-Intensive Research Infrastructure for South Africa [13] commenced in 2015 and will deliver infrastructure broadly equivalent to EUDAT. [↑](#footnote-ref-1)